W81

METHOD AND DEVICE FOR EXTRACTING FEATURE CHARACTER STRING, METHOD AND DEVICE FOR RETRIEVING PSEUDO DOCUMENT USING THEM, STORAGE MEDIUM FOR STORING FEATURE CHARACTER STRING EXTRACTING PROGRAM AND STORAGE MEDIUM FOR STORING PSEUDO DOCUMENT RETRIEVING PROGRAM

Patent Number: JP11338883

Publication date: 1999-12-10

Inventor(s): MATSUBAYASHI TADATAKA; TADA KATSUMI; OKAMOTO TAKUYA; SUGAYA NATSUKO; KAWASHITA YASUSHI

Applicant(s):: HITACHI LTD

Requested Patent: JP11338883

Application Number: JP19980148721 19980529

Priority Number (s):

IPC Classification: G06F17/30 ; G06F17/27

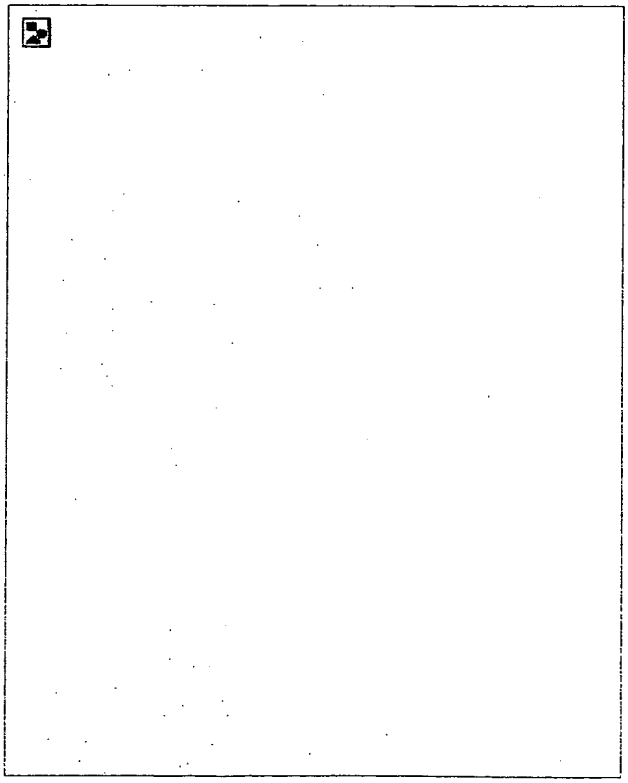
EC Classification:

Equivalents: CN1237738

Abstract

PROBLEM TO BE SOLVED: To provide a method for extracting the feature of contents which are described in a document without using a word dictionary and a high speed pseudo document retrieving system with high precision, where the method is used.

SOLUTION: The system is provided with a step for storing the probability of a character string existing in a text 150 in a text database to appear in the boundary of a word in the text 150 as an appearance probability file 152, a step for storing the times of appearance of the character string existing in the text 150 as an appearance time file 153, a step for extracting a feature character string from the text which is designated by a user through the use of the file 152 and a step for counting the times of the appearance of the feature character string in the text which is designated by the user. Then, a similarity degree as against the text designated by the user is calculated through the use of the file 153 and the number of appearance times in the text which is designated by the user.



W81

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－338883

(43)公開日　平成11年(1999)12月10日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ａ
17/27			15/38	Ｅ
			15/40	３７０Ａ

審査請求　未請求　請求項の数11　ＯＬ　（全 40 頁）

(21)出願番号　特願平10－148721

(22)出願日　平成10年(1998)５月29日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　松林　忠孝
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内

(72)発明者　多田　勝己
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内

(72)発明者　岡本　卓哉
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内

(74)代理人　弁理士　小川　勝男

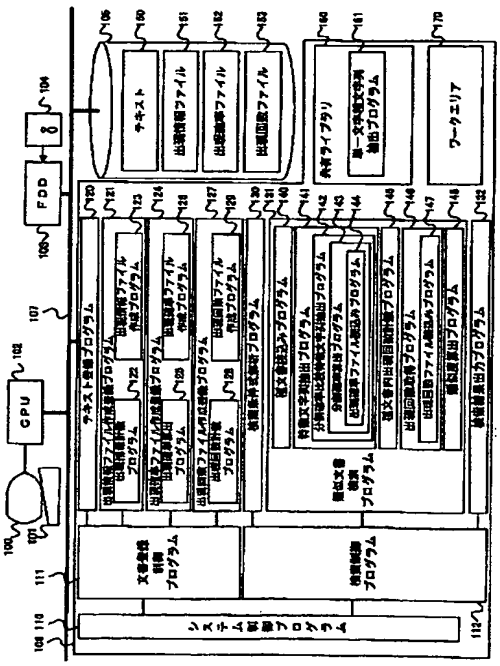
最終頁に続く

(54)【発明の名称】　特徴文字列抽出方法および装置とこれを用いた類似文書検索方法および装置並びに特徴文字列抽出プログラムを格納した記憶媒体および類似文書検索プログラムを格納した記憶媒体

(57)【要約】

【課題】本発明の課題は、単語辞書を用いずに文書に記述された内容の特徴を抽出する方法と、この方法を用いて、高速で高精度な類似文書検索システムを提供することである。

【解決手段】テキストデータベース中のテキスト１５０に存在する文字列のそのテキスト１５０における単語の境界に出現する確率を出現確率ファイル１５２として格納するステップと、テキスト１５０に存在する文字列の出現回数を出現回数ファイル１５３として格納するステップと、出現確率ファイル１５２を用いてユーザが指定したテキストから特徴文字列を抽出するステップと、ユーザが指定したテキストにおける特徴文字列の出現回数を計数するステップとを有し、出現回数ファイル１５３とユーザが指定したテキストにおける出現回数を用いてユーザが指定したテキストに対する類似度を算出する。



【特許請求の範囲】
【請求項１】テキストを含む文書から特徴を表す文字列（特徴文字列と呼ぶ）を抽出する特徴文字列抽出方法において、
単語間の区切れ目を境界として単語の候補となる文字列を上記テキストから抽出する文字列抽出ステップと、
上記文字列抽出ステップで抽出された文字列中の長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出ステップとを有することを特徴とした特徴文字列抽出方法。
【請求項２】請求項１記載の特徴文字列抽出方法における前記文字列抽出ステップとして、
所定の文字種の変わり目を境界としてテキストから単語の候補となる文字列を抽出する文字列抽出ステップを有することを特徴とした特徴文字列抽出方法。
【請求項３】請求項１記載の特徴文字列抽出方法における前記特徴n-gram抽出ステップとして、
前記文字列抽出ステップで単語の候補として抽出された文字列からm-gram（ｍは１以上の整数）とn-gram（ｎは１以上の整数）を抽出し、
該m-gramと該n-gramに関するテキストデータベース内での出現回数を参照し、両者のうち出現回数の多い方を特徴文字列として抽出する特徴n-gram抽出ステップを有することを特徴とした特徴文字列抽出方法。
【請求項４】請求項１記載の特徴文字列抽出方法において、テキストデータベースへの文書登録時の処理として、
テキストから単語の区切れ目を抽出し、これを境界として単語の候補となる文字列を抽出する登録用文字列抽出ステップと、
上記登録用文字列抽出ステップで抽出された文字列（抽出文字列と呼ぶ）に関し、テキストデータベース中での出現回数を計数し、テキストデータベース中の全ての抽出文字列の出現回数に対する割合から出現確率を算出する出現確率算出ステップを有するとともに、
前記特徴n-gram抽出ステップにおいて、出現回数の代わりに該当する出現確率を参照し、前記文字列抽出ステップで抽出された文字列中のn-gramの出現確率を参照し、該出現確率が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出ステップとを有することを特徴とした特徴文字列抽出方法。
【請求項５】文字情報をコードデータとして蓄積したテキストデータベースを対象として、ユーザが指定した文章あるいは文書（以後、まとめて文書と呼ぶ）と類似する文書を検索する類似文書検索方法において、
ユーザが指定した文書のテキスト（指定テキストと呼ぶ）から、単語間の区切れ目を抽出し、これを境界として単語の候補となる文字列を抽出する文字列抽出ステップと、
上記文字列抽出ステップで抽出された文字列の中から、長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出ステップと、
上記特徴n-gram抽出ステップで抽出された特徴文字列に対して、指定テキスト内の出現回数を計数する指定テキスト内出現回数計数ステップと、
上記特徴n-gram抽出ステップで抽出された特徴文字列に対して、テキストデータベース内の各文書における出現回数を取得するテキストデータベース内出現回数取得ステップと、
上記指定テキスト内出現回数計数ステップで計数した該特徴文字列の指定テキスト内の出現回数と、上記テキストデータベース内出現回数取得ステップで取得した該特徴文字列のテキストデータベース内の各文書における出現回数を用いて、指定テキストとテキストデータベース内の各文書の類似度を算出する類似度算出ステップと、
上記類似度算出ステップで算出されたテキストデータベース内の各文書の指定テキストに対する類似度を、検索結果として出力する検索結果出力ステップとを有することを特徴とした類似文書検索方法。
【請求項６】請求項５記載の類似文書検索方法において、テキストデータベースへの文書登録処理として、
テキストから単語の区切れ目を抽出し、これを境界として単語の候補となる文字列を抽出する登録用文字列抽出ステップと、
上記登録用文字列抽出ステップで抽出された文字列から、長さが１から該文字列自体の長さｍまでの全てのn-gramを抽出し、該登録文書の識別番号と該登録文書のテキストにおける出現回数を組みとして、これを該当する出現回数ファイルへ格納する出現回数ファイル作成ステップを有するとともに、
前記テキストデータベース内出現回数取得ステップとして、
前記特徴n-gram抽出ステップで抽出された特徴文字列に対し、該当する上記出現回数ファイルを参照して、該特徴文字列のテキストデータベース内の各文書における該特徴文字列の出現回数を取得するテキストデータベース内出現回数取得ステップを有する類似文書検索方法。
【請求項７】請求項６記載の類似文書検索方法における前記文字列抽出ステップとして、
所定の文字種の変わり目を境界としてテキストから単語の候補となる文字列を抽出する文字列抽出ステップを有することを特徴とした類似文書検索方法。
【請求項８】テキストを含む文書の特徴を表す文字列（特徴文字列と呼ぶ）を抽出する特徴文字列抽出装置において、
単語間の区切れ目を境界として単語の候補となる文字列

を上記テキストから抽出する文字列抽出手段と、
上記文字列抽出装置で抽出された文字列中の長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出手段とを備えたことを特徴とした特徴文字列抽出装置。
【請求項９】文字情報をコードデータとして蓄積したテキストデータベースを対象として、ユーザが指定した文章あるいは文書（以後、まとめて文書と呼ぶ）と類似する文書を検索する類似文書検索装置において、
ユーザが指定した文書のテキスト（指定テキストと呼ぶ）から、単語間の区切れ目を検出し、これを境界として単語の候補となる文字列を抽出する文字列抽出手段と、
上記文字列抽出手段で抽出された文字列の中から、長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出手段と、
上記特徴n-gram抽出手段で抽出された特徴文字列に対して、指定テキスト内の出現回数を計数する指定テキスト内出現回数計数手段と、
上記特徴n-gram抽出手段で抽出された特徴文字列に対して、テキストデータベース内の各文書における出現回数を取得するテキストデータベース内出現回数取得手段と、
上記指定テキスト内出現回数計数ステップで計数した該特徴文字列の指定テキスト内の出現回数と、上記テキストデータベース内出現回数取得手段で取得した該特徴文字列のテキストデータベース内の各文書における出現回数を用いて、指定テキストとテキストデータベース内の各文書の類似度を算出する類似度算出手段と、
上記類似度算出手段で算出したテキストデータベース内の各文書の指定テキストに対する類似度を、検索結果として出力する検索結果出力手段とを備えたことを特徴とした類似文書検索方法。
【請求項１０】テキストを含む文書の特徴を表す文字列（特徴文字列と呼ぶ）を抽出する特徴文字抽出プログラムを格納する記憶媒体において、
単語間の区切れ目を境界として単語の候補となる文字列を上記テキストから抽出する文字列抽出ステップと、
上記文字列抽出ステップで抽出された文字列中の長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出ステップとを有する特徴文字列抽出プログラムを格納することを特徴とした記憶媒体。
【請求項１１】文字情報をコードデータとして蓄積したテキストデータベースを対象として、ユーザが指定した文章あるいは文書（以後、まとめて文書と呼ぶ）と類似する文書を検索する類似文書検索プログラムを格納する記憶媒体において、
ユーザが指定した文書のテキスト（指定テキストと呼ぶ）から、単語間の区切れ目を検出し、これを境界として単語の候補となる文字列を抽出する文字列抽出ステップと、
上記文字列抽出ステップで抽出された文字列の中から、長さがｎ（ｎは１以上の整数）の連続する文字列（n-gramと呼ぶ）に関するテキストデータベース内での出現回数を参照し、該出現回数が最大のn-gramを特徴文字列として抽出する特徴n-gram抽出ステップと、
上記特徴n-gram抽出ステップで抽出された特徴文字列に対して、指定テキスト内の出現回数を計数する指定テキスト内出現回数計数ステップと、
上記特徴n-gram抽出ステップで抽出された特徴文字列に対して、テキストデータベース内の各文書における出現回数を取得するテキストデータベース内出現回数取得ステップと、
上記指定テキスト内出現回数計数ステップで計数した該特徴文字列の指定テキスト内の出現回数と、上記テキストデータベース内出現回数取得ステップで取得した該特徴文字列のテキストデータベース内の各文書における出現回数を用いて、指定テキストとテキストデータベース内の各文書の類似度を算出する類似度算出ステップと、
上記類似度算出ステップで算出されたテキストデータベース内の各文書の指定テキストに対する類似度を、検索結果として出力する検索結果出力ステップを有する類似文書検索プログラムを格納することを特徴とした記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文書に記述された内容の特徴を表す文字列を抽出する方法および装置並びに文字列抽出プログラムを格納した記憶媒体と、この方法および装置を用いて、ユーザが指定した文書に記述されている内容と類似する内容を含む文書を文書データベースの中から検索する方法および装置並びに検索プログラムを格納した記憶媒体に関する。
【０００２】
【従来の技術】近年、パーソナルコンピュータやインターネット等の普及に伴い、電子化文書が爆発的に増加しており、今後も加速度的に増大していくものと予想される。このような状況において、ユーザが所望する情報を含んだ文書を高速かつ効率的に検索したいという要求が高まってきている。
【０００３】このような要求に応える技術として全文検索がある。全文検索では、検索対象文書をテキストとして計算機システムに登録してデータベース化し、この中からユーザが指定した検索文字列（以下、検索タームと

呼ぶ）を含む文書を検索する。このように全文検索では、文書中の文字列そのものを対象として検索を行うため、予めキーワードを付与し、このキーワードを手掛りに検索する従来のキーワード検索システムとは異なり、どんな言葉でも検索ができるという特長がある。

【０００４】しかし、ユーザが所望する情報を含んだ文書を的確に検索するためには、ユーザの検索意図を正確に表わす複雑な検索条件式を作成し、入力する必要がある。これは、情報検索の専門家でない一般のユーザにとっては容易なことではない。

【０００５】この繁雑さを解消するために、ユーザが自分の所望する内容を含んだ文書（以下、種文書と呼ぶ）を例示し、その文書と類似する文書を検索する類似文書検索技術が注目されている。

【０００６】類似文書検索の方法としては、例えば、「特開平８－３３５２２２号公報」に、形態素解析により種文書中に含まれる単語を抽出し、これを用いて類似文書を検索する技術（以下、従来技術１と呼ぶ）が開示されている。

【０００７】従来技術１では、形態素解析により種分書中に含まれる単語を抽出し、この単語を含む文書を類似文書として検索する。例えば、文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」を種文書とする場合、形態素解析により単語辞書を参照して、「携帯電話」「マナー」「問題」等の単語を抽出する。この結果、「携帯電話」を含む文書２「・・・。電車内での携帯電話の使用は禁止されている・・・」を類似文書として検索することができる。

【０００８】しかし、従来技術１では、単語の抽出に単語辞書を用いるため、次のような２つの問題がある。

【０００９】まず、単語辞書に掲載されていない単語が文書の本質的な内容（以下、中心概念と呼ぶ）を表わす場合、この単語が種文書から検索用の単語として抽出されないため、他の単語によって類似検索が行われたとしても、文書の中心概念が正確に検索できない恐れがある。すなわち、ユーザが所望する情報が新語で表されるような場合、これが単語辞書に含まれていないと、目的とする中心概念からずれた文書が検索されてしまうという問題がある。

【００１０】次に、ユーザが所望する情報を表わす言葉が単語辞書に掲載されている場合でも、単語の抽出の仕方によっては検索の対象とする中心概念がずれてしまうという問題がある。例えば、上記の文書１「・・・。携帯電話の使用のマナーが問題になる。・・・」という種文書からは、「携帯電話」「マナー」「問題」等の単語が抽出される。しかし、「電話」という単語が抽出されないため文書３「・・・。電話での話し方について注意された。・・・」という文書の類似度が低く算出されてしまう恐れがある。

【００１１】これらは、全て単語辞書を用いて検索用の単語を抽出する方法を用いていることに起因する。

【００１２】以上が従来技術１の問題点である。

【００１３】この問題を解決するために、「特願平９－３０９０７８号」で、単語辞書を用いずに、種文書中から漢字やカタカナ等の文字種別に連続する n文字の文字列（以下、n-gramと呼ぶ）を漢字やカタカナ等の文字種別に機械的に抽出し、これを用いて類似文書を検索する技術（以下、従来技術２と呼ぶ）を提案した。

【００１４】従来技術２では、文字種別にn-gramの抽出方法を変え、意味のまとまりをもったn-gram（以下、特徴文字列と呼ぶ）を抽出する。例えば、漢字で構成される文字列（以下、漢字文字列と呼ぶ）からは機械的に2-gramを抽出し、カタカナで構成される文字列（以下、カタカナ文字列と呼ぶ）からは、カタカナで構成される最長の文字列（以下、カタカナ最長文字列と呼ぶ）、すなわちカタカナ文字列そのものを抽出する。この場合、上記の文書１「・・・。携帯電話の使用のマナーが問題になる。・・・」という種文書からは、「携帯」「帯電」「電話」「使用」「マナー」「問題」等という特徴文字列が抽出される。すなわち、「電話」という文字列も漏れなく抽出されるため、従来技術１では低い類似度が算出されてしまう文書３「・・・。電話での話し方について注意された。・・・」についても正しく類似度が算出されるようになる。

【００１５】しかし、従来技術２では、複合語を構成する可能性のある漢字文字列等からは、単語間にまたがるn-gramも抽出する可能性がある。このため、これを検索に用いると、内容の類似しない文書に対してまでも類似度が算出され、この結果、関連のない文書が類似文書として検索されるという問題が生じる。例えば、上記の文書１「・・・。携帯電話の使用のマナーが問題になる。・・・」という種文書から抽出された「帯電」という特徴文字列により類似度が算出され、文書４「・・・。電荷の帯電を防ぐために、接地しなくてはならない。・・・」という文書が類似文書として誤って検索されてしまうという問題がある。

【００１６】この問題を解決するための技術として、「情報処理学会論文誌 pp.2286～2297, Vol.38, No.11, Nov.1997」に、1-gramの統計情報を用いて特徴文字列を抽出する技術（以下、従来技術３と呼ぶ）が提案されている。

【００１７】従来技術３では、文書登録時に登録文書中に出現する各1-gramについて、単語の先頭である確率（以下、先頭確率と呼ぶ）と末尾である確率（以下、末尾確率と呼ぶ）を算出しておく。ここでは、単語を、漢字やカタカナ等の文字種境界で区切られ、単一の文字種で構成される文字列（以下、単一文字種文字列と呼ぶ）とし、文字種境界の直後に位置する1-gramを単語の先頭にある1-gramとし、文字種境界の直前に位置する1-gramを単語の末尾にある1-gramとしている。

【００１８】例えば、上記の文書１「・・・。携帯電話の使用のマナーが問題になる。・・・」から文字種境界で抽出した“使用”という漢字文字列では、“使”が単語の先頭にある1-gramで、“用”が単語の末尾にある1-gramとなる。

【００１９】類似文書検索時には、まず指定された種文書から単一文字種文字列を抽出する。次に、単一文字種文字列内の連続する２個の1-gramにおける前方の1-gramの末尾確率と後方の1-gramの先頭確率から、これらの1-gram間で単一文字種文字列が分割される確率（以下、分割確率と呼ぶ）を算出し、この値が所定の値（以下、分割閾値と呼ぶ）を越えている場合には、そこで単一文字種文字列を分割するという処理を行う。

【００２０】以下、分割閾値を0.050として、従来技術３の具体的な処理方法を説明する。

【００２１】まず、文書登録時には全登録対象文書中に出現する各1-gramについて、出現回数、単語の先頭に出現する回数（以下、先頭回数と呼ぶ）および末尾に出現する回数（以下、末尾回数と呼ぶ）を計数し、出現情報ファイルに格納する。例えば、上記の文書１では“携”の出現回数は１回、先頭回数は１回および末尾回数は０回という出現情報が得られる。図２に出現情報ファイルの例を示す。

【００２２】その後、上記出現情報ファイルを参照し、各1-gramについて、それぞれ先頭確率と末尾確率を算出し、出現確率ファイルに格納する。例えば、1-gram“携”の先頭確率は768 / 4,740 = 0.16、末尾確率は 492 / 4,740 = 0.10となる。図３に出現確率ファイルの例を示す。

【００２３】次に、単一文字種文字列「携帯電話」を例として、従来技術３の文書検索方法を説明する。

【００２４】まず、単一文字種文字列「携帯電話」の中から1-gramの二つの組として、（“携”，“帯”）、（“帯”，“電”）および（“電”，“話”）の３個を抽出する。次に、各1-gramの組において、前方の1-gramの末尾確率と後方の1-gramの先頭確率を、登録時に作成した出現確率ファイルから取得し、分割確率を算出する。

【００２５】図４に、「携帯電話」から抽出した３個の1-gramの組における分割確率の算出過程を示す。本例では、（“携”，“帯”）、（“帯”，“電”）および（“電”，“話”）の分割確率として、それぞれ0.011、0.054および0.005が算出され、これらの分割確率のうち、（“帯”，“電”）の0.054が分割閾値0.050より大きいので、“帯”と“電”の間で分割される。一方、（“携”，“帯”）および（“電”，“話”）の分割確率はそれぞれ0.011および0.005であり、これらは分割閾値0.050より小さいので、これらの1-gram間では分割されない。その結果、「携帯電話」が“帯”と“電”の間で分割され、「携帯」と「電話」の２個の特徴文字列が抽出されることになる。

【００２６】以上が、従来技術３の具体的な処理方法である。このように従来技術３では、1-gramの統計情報を用いて特徴文字列を抽出することにより、単語間にまたがる不適切な特徴文字列を抽出しないようにして、内容の類似しない文書が検索されることのないように配慮している。

【００２７】しかし、従来技術３では、分割確率の絶対値で分割の可否を判断するため、単語としての特徴文字列の抽出精度が低いという問題がある。例えば、単一文字種文字列「帯電」に対しては、1-gramの組（“帯”，“電”）が抽出され、この1-gram間の分割確率として0.054が算出される。

【００２８】この値は分割閾値0.050より大きいため、「帯電」が「帯」と「電」のように誤って分割（以下、誤分割と呼ぶ）されてしまい、不適切な２個の特徴文字列が抽出されてしまう。この結果、「帯（おび）」に関係のある文書等も類似文書として検索されてしまい、検索ノイズが混入して、目的とする中心概念がずれた文書が類似文書として検索されてしまうという問題がある。

【００２９】

【発明が解決しようとする課題】以上述べたように、従来技術１のように単語辞書を用いて単語を抽出する方法では、単語辞書に掲載されていない単語が種文書の中心概念を表す場合には、中心概念からずれた文書が検索されてしまうという問題がある。

【００３０】また、従来技術２のように単一文字種文字列から文字種別に、単純にn-gramを抽出する方法では、複合語を構成する可能性のある漢字文字列等から単語間にまたがるn-gramを抽出してしまうことにより、関連のない文書が類似文書として検索されてしまうという問題がある。

【００３１】さらに、従来技術３のように、1-gramの統計情報を用いて分割確率を算出し、この値の絶対値で分割の可否を判断する方法においても、単語としての特徴文字列の抽出精度が低いため、検索ノイズが混入し、目的とする中心概念がずれた文書が類似文書として検索されてしまうという問題がある。

【００３２】こうした従来技術の問題に対し、本発明では、誤分割が少なくなるように特徴文字列を抽出する方法および装置を提供することを目的とする。

【００３３】また、誤分割が少なくなるように特徴文字列を抽出することにより、検索ノイズを少なくすることで中心概念のずれを低減した類似文書検索が行える方法および装置を提供することを目的とする。

【００３４】

【課題を解決するための手段】上記課題を解決するために、本発明による特徴文字列抽出方法では、以下に示すステップからなる処理により、種文書から特徴文字列の抽出を行なう。

【００３５】すなわち、本発明による特徴文字列抽出方法では、文書の登録処理として、
（ステップ１）登録対象文書を読み込む文書読込みステップ、
（ステップ２）上記文書読込みステップで読み込んだ登録対象文書中の文字列を、漢字やカタカナ等の文字種境界で分割し、単一文字種文字列として抽出する単一文字種文字列抽出ステップ、
（ステップ３）上記単一文字種文字列抽出ステップで抽出された単一文字種文字列に対して、その文字種を判定し、漢字やカタカナならば予め定められた長さのn-gramについて登録文書における出現回数、単語の先頭に出現する回数（以下、先頭回数と呼ぶ）と末尾に出現する回数（以下、末尾回数と呼ぶ）、およびn-gramそのものが単語として出現する回数（以下、単独回数と呼ぶ）を計数する出現情報計数ステップ、
（ステップ４）上記出現情報計数ステップで計数されたn-gramの出現情報を、既にデータベースに登録されている文書に関する該n-gramの出現情報に加算することで、データベース全体の出現情報を算出し、該当する出現情報ファイルへ格納する出現情報ファイル作成登録ステップ、
（ステップ５）上記出現情報計数ステップで出現情報が計数されたn-gramに関して、該当する出現情報ファイルからデータベース全体における出現情報を取得し、単語の先頭である確率（以下、先頭確率と呼ぶ）と末尾である確率（以下、末尾確率と呼ぶ）およびn-gramそのものが単語として出現する確率（以下、単独確率と呼ぶ）を算出し、該当する出現確率ファイルに格納する出現確率ファイル作成登録ステップ、
（ステップ６）上記単一文字種文字列抽出ステップで抽出された単一文字種文字列から、予め定められた長さのn-gramを抽出し、登録対象文書中における出現回数を計数する出現回数計数ステップ、
（ステップ７）上記出現回数計数ステップで計数された出現回数を該当する出現回数ファイルに格納する出現回数ファイル作成登録ステップ、を有し、種文書から特徴文字列を抽出する処理として、
（ステップ８）種文書を読み込む種文書読込みステップ、
（ステップ９）上記種文書読込みステップにおいて読み込まれた種文書中の文字列を文字種境界で分割し、単一文字種文字列として抽出する検索用単一文字種文字列抽出ステップ、
（ステップ１０）上記検索用単一文字種文字列抽出ステップで抽出された単一文字種文字列に関して、その文字種を判定し、漢字やカタカナならば、前記出現確率ファイルを読み込み、単一文字種文字列の先頭から i文字目までの文字列の単独確率、(i+1)文字目までの文字列の単独確率、(i+1)文字目の文字の先頭確率および(i+2)文字目の文字の先頭確率を取得し、i文字目で単一文字種文字列が分割される確率（以下、分割確率と呼ぶ）を i文字目までの文字列の単独確率と(i+1)文字目の文字の先頭確率の積として算出し、（ｉ＋１）文字目での分割確率を、(i+1)文字目までの文字列の単独確率と(i+2)文字目の文字の先頭確率の積として算出し、これらの i文字目と(i+1)文字目の分割確率を比較して、値の大きい方を単一文字種文字列が分割される点（以下、分割点と呼ぶ）とし、先頭から分割点までの文字列を特徴文字列として抽出し、漢字やカタカナ以外ならば、単一文字種文字列そのものを特徴文字列として抽出し、抽出された特徴文字列を除外した残りの文字列に対して、同様の処理を繰り返すことによって特徴文字列を抽出する特徴文字列抽出ステップを有する。
【００３６】また、前述の課題を解決するために、本発明による類似文書検索方法では、上記ステップからなる処理により、種文書と類似する文書を検索するための特徴文字列を抽出し、これを用いて類似文書検索を行う。
【００３７】すなわち、本発明による類似文書検索方法では、文書の登録処理として、
（ステップ１）登録対象文書を読み込む文書読込みステップ、
（ステップ２）上記文書読込みステップで読み込んだ登録対象文書中の文字列を、漢字やカタカナ等の文字種境界で分割し、単一文字種文字列として抽出する単一文字種文字列抽出ステップ、
（ステップ３）上記単一文字種文字列抽出ステップで抽出された単一文字種文字列に対して、その文字種を判定し、漢字やカタカナならば予め定められた長さのn-gramについて登録文書における出現回数、単語の先頭に出現する回数（以下、先頭回数と呼ぶ）と末尾に出現する回数（以下、末尾回数と呼ぶ）、およびn-gramそのものが単語として出現する回数（以下、単独回数と呼ぶ）を計数する出現情報計数ステップ、
（ステップ４）上記出現情報計数ステップで計数されたn-gramの出現情報を、既にデータベースに登録されている文書に関する該n-gramの出現情報に加算することで、データベース全体の出現情報を算出し、該当する出現情報ファイルへ格納する出現情報ファイル作成登録ステップ、
（ステップ５）上記出現情報計数ステップで出現情報が計数されたn-gramに関して、該当する出現情報ファイルからデータベース全体における出現情報を取得し、単語の先頭である確率（以下、先頭確率と呼ぶ）と末尾である確率（以下、末尾確率と呼ぶ）およびn-gramそのものが単語として出現する確率（以下、単独確率と呼ぶ）を算出し、該当する出現確率ファイルに格納する出現確率ファイル作成登録ステップ、
（ステップ６）上記単一文字種文字列抽出ステップで抽出された単一文字種文字列から、予め定められた長さの

n-gramを抽出し、登録対象文書中における出現回数を計数する出現回数計数ステップ、
（ステップ７）上記出現回数計数ステップで計数された出現回数を該当する出現回数ファイルに格納する出現回数ファイル作成登録ステップ、を有し、種文書に類似する文書の検索処理として、
（ステップ８）種文書を読み込む種文書読込みステップ、
（ステップ９）上記種文書読込みステップにおいて読み込まれた種文書中の文字列を文字種境界で分割し、単一文字種文字列として抽出する検索用単一文字種文字列抽出ステップ、
（ステップ１０）上記検索用単一文字種文字列抽出ステップで抽出された単一文字種文字列に関して、その文字種を判定し、漢字やカタカナならば、前記出現確率ファイルを読み込み、単一文字種文字列の先頭から i文字目までの文字列の単独確率、(i+1)文字目までの文字列の単独確率、(i+1)文字目の文字の先頭確率および(i+2)文字目の文字の先頭確率を取得し、i文字目で単一文字種文字列が分割される確率（以下、分割確率と呼ぶ）を i文字目までの文字列の単独確率と(i+1)文字目の文字の先頭確率の積として算出し、(i+1)文字目での分割確率を、(i+1)文字目までの文字列の単独確率と(i+2)文字目の文字の先頭確率の積として算出し、これらの i文字目と(i+1)文字目の分割確率を比較して、値の大きい方を単一文字種文字列が分割される点（以下、分割点と呼ぶ）とし、先頭から分割点までの文字列を特徴文字列として抽出し、漢字やカタカナ以外ならば、単一文字種文字列そのものを特徴文字列として抽出し、抽出された特徴文字列を除外した残りの文字列に対して、同様の処理を繰り返すことによって特徴文字列を抽出する特徴文字列抽出ステップ、
（ステップ１１）上記特徴文字列抽出ステップで抽出された全ての特徴文字列に対して、種文書内における出現回数を計数する種文書内出現回数計数ステップ、
（ステップ１２）上記特徴文字列抽出ステップで抽出された全ての特徴文字列に対して、前記出現回数ファイルを読み込み、データベース内の各文書における該当特徴文字列の出現回数を取得するデータベース内出現回数取得ステップ、
（ステップ１３）上記特徴文字列抽出ステップで抽出された特徴文字列に対し、上記種文書内出現回数計数ステップで計数された種文書内の出現回数と、上記データベース内出現回数取得ステップで取得されたデータベース内の各文書における出現回数を用いて、予め定められた算出式に基づいて種文書とデータベース内の各文書との類似度を算出する類似度算出ステップ、
（ステップ１４）上記類似度算出ステップで算出された類似度に基づいて、検索結果を出力する検索結果出力ステップを有する。

【００３８】上記文書検索方法を用いた本発明の原理を、以下に説明する。
【００３９】本発明では、文書を登録する際に、（ステップ１）～（ステップ７）を実行する。
【００４０】まず、文書読込みステップ（ステップ１）で登録対象となる文書を読み込む。次に、単一文字種文字列抽出ステップ（ステップ２）において、上記文書読込みステップ（ステップ１）で読み込まれた登録対象文書中の文字列を、漢字やカタカナ等の文字種境界で分割し、単一文字種からなる文字列を抽出する。例えば、前述の文書２「・・・。電車内での携帯電話の使用は禁止されている。・・・」という文書からは、「電車内」「での」「携帯電話」「の」「使用」「は」「禁止」「されている」等の単一文字種文字列が抽出される。
【００４１】次に、出現情報計数ステップ（ステップ３）において、単一文字種文字列抽出ステップ（ステップ２）で抽出された上記各単一文字種文字列について、その文字種を判定し、漢字やカタカナならば予め定められた長さ nのn-gramの登録対象文書中の出現回数、先頭回数、末尾回数および単独回数を計数する。例えば、漢字文字列とカタカナ文字列から1-gramおよび2-gramの出現回数、先頭回数および末尾回数を計数するものと定められている場合には、上記単一文字種文字列抽出ステップ（ステップ２）で抽出された単一文字種文字列について、“携”の出現回数は１回、そのうち先頭回数は１回、末尾回数は０回、単独回数は０回であり、“携帯”の出現回数は１回、そのうち先頭回数は１回、末尾回数は０回、単独回数は０回と計数される。
【００４２】次に、出現情報ファイル作成登録ステップ（ステップ４）において、先に出現情報計数ステップ（ステップ３）で抽出されたn-gramの出現情報を、既にデータベースに登録されている文書に関する出現情報に加算し、累積情報としての出現情報を該当する出現情報ファイルへ格納する。図５に出現情報ファイルの例を示す。本図に示した出現情報ファイルは、上記出現情報計数ステップ（ステップ３）において抽出された出現情報を格納した場合の例である。本図に示した出現情報ファイルは、前述の1-gram“携”に関しては、出現回数4,740回、先頭回数768回、末尾回数492回、および単独回数42回という情報を格納し、2-gram“携帯”に関しては、出現回数462回、先頭回数419回、末尾回数52回、および単独回数48回という情報を格納していることを表わす。
【００４３】次に、出現確率ファイル作成登録ステップ（ステップ５）において、出現情報ファイル作成登録ステップ（ステップ４）で出現情報が格納されたn-gramに対して、それぞれ出現確率を算出し、該当する出現確率ファイルに格納する。例えば、図５に示すように、1-gram“携”に関しては、出現回数4,740回、先頭回数768回、末尾回数492回、および単独回数42回であることから、先頭確率は768 / 4,740 = 0.16、末尾確率は 492 /

4,740 = 0.10、単独確率は42 / 4,740 = 0.01と計算される。図６に出現確率ファイルの例を示す。本図に示した出現確率ファイルは、上記出現情報計数ステップ（ステップ３）において抽出された出現確率を格納した場合の例であり、前述の1-gram“携”に関しては、先頭確率0.16、末尾確率0.10、および単独確率0.01という情報が格納され、2-gram“携帯”に関しては、先頭確率0.90、末尾確率0.11、および単独確率0.10という情報が格納されていることを表わす。

【００４４】次に、出現回数計数ステップ（ステップ６）において、単一文字種文字列抽出ステップ（ステップ２）で抽出された全ての単一文字種文字列から、予め定められた長さのn-gramを抽出し、登録対象文書中における出現回数を計数する。そして、出現回数ファイル作成登録ステップ（ステップ７）において、上記出現回数計数ステップ（ステップ６）で抽出された各n-gramの出現回数を該当する出現回数ファイルに格納する。

【００４５】図２４に、前述の文書２「・・・。電車内での携帯電話の使用は禁止されている。・・・」を例に、出現回数ファイル作成処理の手順を示す。

【００４６】まず、単一文字種文字列抽出ステップ（ステップ２）で登録対象文書である文書２から全ての単一文字種文字列を抽出する。

【００４７】次に、出現回数計数ステップ（ステップ６）で、上記単一文字種文字列抽出ステップ（ステップ２）で抽出された全ての単一文字種文字列から予め定められた長さのn-gramを抽出し、登録対象文書内の出現回数を計数する。本図に示した例では、単一文字種文字列から長さが３のn-gramまでを抽出するものとし、単一文字種文字列２４０４に含まれる「電車内」から、長さが１の“電”、“車”、“内”、長さが２の“電車”、“車内”、および長さが３の“電車内”が抽出され、文書２における出現回数が計数される。この結果、“電”は文書２の中に２回出現し、“車”は文書２の中に１回出現しているというように計数される。

【００４８】そして、出現回数ファイル作成登録ステップ（ステップ７）で、出現回数計数ステップ（ステップ６）で抽出された各n-gramの出現回数を該当する出現回数ファイルに格納する。この結果、文書２からは、1-gram“電”（２，２）、“車”（２，１）、“内”（２，１）、2-gram“電車”（２，１）、“車内”（２，１）、3-gram“電車内”（２，１）のように各n-gramの登録対象文書の識別番号と出現回数が組みとして格納される。ここで、“電車”（２，１）は、2-gram“電車”が文書番号２の文書に、１回出現するということを示している。

【００４９】検索時には、（ステップ８）～（ステップ１４）を実行する。

【００５０】まず、種文書読込みステップ（ステップ８）において、種文書として文書１を読み込む。　次に、検索用単一文字種文字列抽出ステップ（ステップ９）において、上記種文書読込みステップ（ステップ８）で読み込まれた種文書（文書１）中の文字列を文字種境界で分割し、単一文字種文字列を抽出する。

【００５１】次に、特徴文字列抽出ステップ（ステップ１０）において、上記検索用単一文字種文字列抽出ステップ（ステップ９）で抽出された単一文字種文字列について、その文字種を判定する。

【００５２】この文字種が、漢字やカタカナならば、前述した出現確率ファイルを読み込み、単一文字種文字列の先頭から i文字目までの文字列の単独確率、(i+1)文字目までの文字列の単独確率、(i+1)文字目の文字の先頭確率および(i+2)文字目の文字の先頭確率を取得する。そして、 i文字目での分割確率を i文字目までの文字列の単独確率と(i+1)文字目の文字の先頭確率の積として算出し、(i+1)文字目での分割確率を(i+1)文字目までの文字列の単独確率と(i+2)文字目の文字の先頭確率の積として算出する。そして、これらの i文字目と(i+1)文字目の分割確率を比較して、値の大きい方を分割点とし、先頭から該分割点までの文字列を特徴文字列として抽出する。

【００５３】また、漢字やカタカナでなければ、単一文字種文字列そのものを特徴文字列として抽出し、以下、同様の処理を繰り返すことによって、特徴文字列を抽出する。

【００５４】図８に、文書１から抽出した単一文字種文字列「携帯電話」から特徴文字列を抽出する例を示す。まず、「携帯電話」における1文字目での分割確率は、「携」の単独確率0.01と「帯」の先頭確率0.11の積として0.001が算出され、2文字目での分割確率は、「携帯」の単独確率0.10と「電」の先頭確率0.36の積として0.036が算出される。次に、これらの分割確率を比較し、値の大きい方で単一文字種文字列を分割する。この場合、2文字目の分割確率0.036の方が大きいので、単一文字種文字列「携帯電話」は「携帯」と「電話」に分割される。

【００５５】また、図９に、従来技術３では適切に分割されない単一文字種文字列「帯電」の例について、本発明の分割処理を示す。まず、「帯電」における 1文字目での分割確率は、「帯」の単独確率0.01と「電」の単独確率0.01の積として0.0001と算出される。また、 2文字目での分割確率、すなわち「帯電」が単一文字種文字列そのものとして出現する確率は、「帯電」の単独確率0.10と算出される。これらの値を比較して、値の大きい方で単一文字種文字列に分割される。この場合、「帯電」の単独確率0.10の方が大きいので、「帯電」は 2文字目で分割されることになり、結果的に単一文字種文字列「帯電」は分割されず、一塊の文字列として抽出されることになる。

【００５６】このように分割確率を比較して単一文字種

文字列を分割することにより、データベース中での実際の出現状況を正確に反映した単語分割が行なえるため、分割確率の絶対値で分割する前述した従来技術３に比べ、不適切な分割を大幅に削減することが可能になる。

【００５７】次に、種文書内出現回数計数ステップ（ステップ１１）において、上記特徴文字列抽出ステップ（ステップ１０）で抽出された特徴文字列の種文書内における出現回数を計数する。

【００５８】そして、データベース内出現回数取得ステップ（ステップ１２）において、上記特徴文字列抽出ステップ（ステップ１０）で抽出された特徴文字列に対して、前述した出現回数ファイルを参照し、データベース内の各文書における出現回数を得る。

【００５９】そして、類似度算出ステップ（ステップ１３）において、前記特徴文字列抽出ステップ（ステップ１０）で抽出された特徴文字列に対して、上記種文書内出現回数計数ステップ（ステップ１１）とデータベース内出現回数取得ステップ（ステップ１２）で計数された種文書内における出現回数と、データベース内の各文書における出現回数を基に、類似度が算出される。

【００６０】類似度の算出には、例えば、「特開平６－１１０９４８号公報」に開示されている以下に示す類似度算出式（１）を用いてもよい。

【００６１】

【数式１】

$$S(i)=\frac{\sum_{j=1}^{n}(U(j)\times R(j))}{\sqrt{\sum_{j=1}^{n}U(j)^2\times\sum_{i=1}^{n}R(j)^2}} \qquad \cdots (1)$$

【００６２】ここで、U(j)は種文書中の j番目のn-gramの正規化ウエイトを示し、各n-gramの種文書内出現回数から算出される。R(j)はデータベース中文書の j番目のn-gramの正規化ウエイトを示、各n-gramのデータベース内の各文書における出現回数から算出される。正規化ウエイトとは、データベースにおけるn-gramの出現偏りを表し、この値が大きいn-gramほどある特定の文書に偏って出現することを意味する。この正規化ウエイトの算出方法については、「特開平６－１１０９４８号公報」で説明されているため、ここでは説明を省略する。また、nはデータベース中の全文書数を表わす。

【００６３】この類似度算出式（１）を用いて、文書１が種文書として指定された場合の文書 iの類似度S(i)を算出すると、次のようになる。

【００６４】S(1) = 1.0

S(2) = 0.262

S(3) = 0.048

S(4) = 0.0

この結果、検索結果出力ステップ（ステップ１４）で、文書を類似度の降順に整列すると、文書１、文書２、および文書３の順に表示されることになる。類似度が 0の文書４は検索結果としては出力されない。

【００６５】以上説明したように、本発明の特徴文字列抽出方法を用いた類似文書検索方法によれば、 従来技術１のように単語辞書を用いることなく単一文字種文字列から文字列を機械的に抽出することができるため、どのような単語についても漏れなく検索に供することができ、種文書が表わす概念を正確に検索することが可能となる。

【００６６】また、従来技術２のように単一文字種文字列から文字種別に、単純にn-gramを抽出するのではなく、統計情報を用いて意味のまとまったn-gramを抽出することにより、種文書が表わす概念をより正確に検索することが可能となる。

【００６７】さらに、従来技術３のように分割確率の絶対値で分割するのではなく、分割確率を比較し、その値が大きい方で分割することにより、データベース中での実際の出現状況を正確に反映した単語分割が可能となり、不適切な単語分割を大幅に削減することが可能となる。そのため、従来技術３に比べ不適切な特徴文字列が検索に供されないため、種文書が表わす概念を適切に検索できるとともに、高速に類似文書を検索することができるようになる。

【００６８】

【発明の実施の形態】以下、本発明の第一の実施例について図１を用いて説明する。

【００６９】本発明を適用した類似文書検索システムの第一の実施例は、ディスプレイ１００、キーボード１０１、中央演算処理装置（ＣＰＵ）１０２、磁気ディスク装置１０５、フロッピディスクドライブ（ＦＤＤ）１０３、主メモリ１０６およびこれらを結ぶバス１０７から構成される。

【００７０】磁気ディスク装置１０５には、テキスト１５０、出現情報ファイル１５１、出現確率ファイル１５２および出現回数ファイル１５３が格納される。ＦＤＤ１０３を介してフロッピディスク１０４に格納されている登録文書や種文書等の情報が、主メモリ１０６内に確保されるワークエリア１７０あるいは磁気ディスク装置１０５へ読み込まれる。

【００７１】主メモリ１０６には、システム制御プログラム１１０、文書登録制御プログラム１１１、共有ライブラリ１６０、テキスト登録プログラム１２０、出現情報ファイル作成登録プログラム１２１、出現確率ファイル作成登録プログラム１２４、出現回数ファイル作成登録プログラム１２７、検索制御プログラム１１２、検索条件式解析プログラム１３０、類似文書検索プログラム１３１および検索結果出力プログラム１３２が格納されるとともにワークエリア１７０が確保される。これらのプログラムは、フロッピーディスクやＣＤ－ＲＯＭなどの持ち運び可能な記憶媒体に格納され、ここから読み出

し磁気ディスク装置１０５へインストールする。本装置起動時に、システム制御プログラム１１０が起動し、これらのプログラムを磁気ディスク装置１０５から読み出し、主メモリ１０６へ格納する。

【００７２】共有ライブラリ１６０は、単一文字種文字列抽出プログラム１６１で構成される。

【００７３】出現情報ファイル作成登録プログラム１２１は、出現情報計数プログラム１２２と出現情報ファイル作成プログラム１２３で構成されるとともに、後述するように共有ライブラリ１６０から単一文字種文字列抽出プログラム１６１を呼び出す構成をとる。

【００７４】出現確率ファイル作成登録プログラム１２４は、出現確率算出プログラム１２５と出現確率ファイル作成プログラム１２６で構成される。

【００７５】出現回数ファイル作成登録プログラム１２７は、出現回数計数プログラム１２８と出現回数ファイル作成プログラム１２９で構成される。

【００７６】類似文書検索プログラム１３１は、種文書読込みプログラム１４０、特徴文字列抽出プログラム１４１、種文書内出現回数計数プログラム１４５、出現回数取得プログラム１４６および類似度算出プログラム１４８で構成されるとともに、後述するように共有ライブラリ１６０から単一文字種文字列抽出プログラム１６１を呼び出す構成をとる。

【００７７】特徴文字列抽出プログラム１４１は、分割確率比較特徴文字列抽出プログラム１４２を呼び出す構成をとる。分割確率比較特徴文字列抽出プログラム１４２は、分割確率算出プログラム１４３を呼び出す構成をとる。分割確率算出プログラム１４３は出現確率ファイル読込みプログラム１４４を呼び出す構成をとる。

【００７８】出現回数取得プログラム１４６は、出現回数ファイル読込みプログラム１４７を呼び出す構成をとる。

【００７９】文書登録制御プログラム１１１および検索制御プログラム１１２は、ユーザによるキーボード１０１からの指示に応じてシステム制御プログラム１１０によって起動され、それぞれテキスト登録プログラム１２０、出現情報ファイル作成登録プログラム１２１、出現確率ファイル作成登録プログラム１２４および出現回数ファイル作成登録プログラム１２７の制御と、検索条件式解析プログラム１３０、類似文書検索プログラム１３１および検索結果出力プログラム１３２の制御を行なう。

【００８０】以下、本実施例における類似文書検索システムの処理手順について説明する。

【００８１】まず、システム制御プログラム１１０の処理手順について図１０のＰＡＤ（Ｐｒｏｂｌｅｍ　Ａｎａｌｙｓｉｓ　Ｄｉａｇｒａｍ）図を用いて説明する。

【００８２】システム制御プログラム１１０では、まずステップ１０００で、キーボード１０１から入力されたコマンドを解析する。

【００８３】次に、ステップ１００１で、この解析結果が登録実行のコマンドであると判定された場合には、ステップ１００２で文書登録制御プログラム１１１を起動して、文書の登録を行なう。

【００８４】またステップ１００３で、検索実行のコマンドであると判定された場合には、ステップ１００４で検索制御プログラム１１２を起動して、類似文書の検索を行なう。

【００８５】以上が、システム制御プログラム１１０の処理手順である。

【００８６】次に、図１０に示したステップ１００２でシステム制御プログラム１１０により起動される文書登録制御プログラム１１１の処理手順について、図１１のＰＡＤ図を用いて説明する。

【００８７】文書登録制御プログラム１１１では、まずステップ１１００でテキスト登録プログラム１２０を起動し、ＦＤＤ１０３に挿入されたフロッピディスク１０４から登録すべき文書のテキストデータをワークエリア１７０に読み込み、これをテキスト１５０として磁気ディスク装置１０５に格納する。テキストデータは、フロッピディスク１０４を用いて入力するだけに限らず、通信回線やＣＤ－ＲＯＭ装置（図１には示していない）等を用いて他の装置から入力するような構成を取ることも可能である。

【００８８】次に、ステップ１１０１で出現情報ファイル作成登録プログラム１２１を起動し、ワークエリア１７０に格納されているテキスト１５０を読み出し、その中の各n-gramに対する出現情報ファイル１５１を作成し、磁気ディスク装置１０５に格納する。

【００８９】次に、ステップ１１０２で出現確率ファイル作成登録プログラム１２４を起動し、ワークエリア１７０に格納されているテキスト１５０中の各n-gramに対する出現確率を算出し、該当する出現確率ファイル１５２として、磁気ディスク装置１０５へ格納する。

【００９０】次に、ステップ１１０３で出現回数ファイル作成登録プログラム１２７を起動し、ワークエリア１７０に格納されているテキスト１５０を読み出し、その中の各文書における全てのn-gramに対する出現回数を計数し、該当する出現回数ファイル１５３として、磁気ディスク装置１０５へ格納する。

【００９１】以上が、文書登録制御プログラム１１１の処理手順である。

【００９２】次に、図１１に示したステップ１１０１で文書登録制御プログラム１１１により起動される出現情報ファイル作成登録プログラム１２１の処理手順について、図１２のＰＡＤ図を用いて説明する。

【００９３】出現情報ファイル作成登録プログラム１２１では、まずステップ１２００で単一文字種文字列抽出プログラム１６１を起動し、テキスト１５０の文字列を

文字種境界で分割することにより単一文字種文字列を抽出し、ワークエリア１７０に格納する。

【００９４】次に、ステップ１２０１において、出現情報計数プログラム１２２を起動し、テキスト１５０における予め定められた長さのn-gramの出現回数と、ワークエリア１７０に格納されている単一文字種文字列の先頭回数、末尾回数および単独回数を計数し、同じくワークエリア１７０に格納する。

【００９５】そして、ステップ１２０２において、出現情報ファイル作成プログラム１２３を起動し、ワークエリア１７０に格納されているテキスト１５０におけるn-gramの出現回数、先頭回数、末尾回数および単独回数を、それぞれ出現情報ファイル１５１に格納されている該当n-gramの出現回数、先頭回数、末尾回数および単独回数に加算し、ワークエリア１７０に格納するとともに出現情報ファイル１５１として磁気ディスク装置１０５に格納する。

【００９６】以上が、出現情報ファイル作成登録プログラム１２１の処理手順である。

【００９７】次に、図１１に示したステップ１１０２で文書登録制御プログラム１１１により起動される出現確率ファイル作成登録プログラム１２４の処理手順について、図１６のＰＡＤ図を用いて説明する。

【００９８】出現確率ファイル作成登録プログラム１２４では、まずステップ１６００で出現確率算出プログラム１２５を起動し、ワークエリア１７０に格納されている各n-gramの出現情報から、各n-gramの単独確率、先頭確率および末尾確率を算出し、ワークエリア１７０へ格納する

次に、ステップ１６０１において、出現確率ファイル作成プログラム１２６を起動し、ワークエリア１７０に格納されている各n-gramの単独確率、先頭確率および末尾確率を出現確率ファイル１５２として磁気ディスク装置１０５に格納する。

【００９９】以上が、出現確率ファイル作成登録プログラム１２４の処理手順である。

【０１００】次に、図１１に示したステップ１１０３で文書登録制御プログラム１１１により起動される出現回数ファイル作成登録プログラム１２７の処理手順について、図２５に示すＰＡＤ図を用いて説明する。

【０１０１】出現回数ファイル作成登録プログラム１２７では、まずステップ２５００で出現回数計数プログラム１２８を起動し、図１２のステップ１２００でワークエリア１７０に格納した全ての単一文字種文字列の中から、長さが１から単一文字種文字列自体の長さｍまでのn-gramを抽出し、登録対象文書におけるそれらの出現回数を計数し、ワークエリア１７０に格納する。

【０１０２】次に、ステップ２５０１において、出現回数ファイル作成プログラム１２９を起動し、ステップ２５００で計数した各n-gramの出現回数を登録対象文書の識別番号（以下、文書番号と呼ぶ）とともに出現回数ファイル１５３として磁気ディスク装置１０５に格納する

次に、図１０に示したステップ１００４でシステム制御プログラム１１０により起動される検索制御プログラム１１２による類似文書検索の処理手順について、図１３のＰＡＤ図を用いて説明する。

【０１０３】検索制御プログラム１１２では、まずステップ１３００で検索条件式解析プログラム１３０を起動し、キーボード１０１から入力された検索条件式を解析し、検索条件式のパラメータとして指定された種文書の文書番号を抽出する。

【０１０４】次に、ステップ１３０１で類似文書検索プログラム１３１を起動し、上記検索条件式解析プログラム１３０により抽出された文書番号の種文書に対し、磁気ディスク装置１０５に格納されているテキスト１５０中の各文書の類似度を算出する。

【０１０５】最後に、ステップ１３０２において、検索結果出力プログラム１３２を起動し、上記類似文書検索プログラム１３１で算出された各文書の類似度に基づいて、検索結果を出力する。

【０１０６】以上が、検索制御プログラム１１２による文書検索の処理手順である。

【０１０７】次に、図１３に示したステップ１３０１で検索制御プログラム１１２により起動される類似文書検索プログラム１３１の処理手順について、図１４のＰＡＤ図を用いて説明する。

【０１０８】類似文書検索プログラム１３１では、まずステップ１４００で種文書読込みプログラム１４０を起動し、検索条件式解析プログラム１３０によって検索条件式から抽出された文書番号の種文書を磁気ディスク装置１０５中のテキスト１５０からワークエリア１７０に読み込む。

【０１０９】ここで、種文書は、テキスト１５０中に格納されている文書を読み込むだけでなく、キーボード１０１から直接入力することも可能であり、フロッピディスク１０４、ＣＤ－ＲＯＭ装置（図１には示していない）や通信回線等を用いて、他の装置から入力するような構成を取ることも可能であり、また、全文検索システム等による検索結果から入力するような構成を取ることも可能であり、さらには、検索結果出力プログラム１３２の出力から種文書を選択する構成を取ることも可能である。

【０１１０】次に、ステップ１４０１において、共有ライブラリ１６０の単一文字種文字列抽出プログラム１６１を起動し、上記種文書読込みプログラム１４０で読み込んだ種文書のテキストを、文字種境界で分割して単一文字種文字列を取得し、ワークエリア１７０に格納する。

【０１１１】そして、ステップ１４０２において、後述する特徴文字列抽出プログラム１４１を起動し、上記単

一文字種文字列抽出プログラム１６１で取得した単一文字種文字列から、特徴文字列を抽出する。

【０１１２】次に、ステップ１４０３において、種文書内出現回数計数プログラム１４５を起動し、上記特徴文字列抽出プログラム１４１で取得した特徴文字列の、種文書内での出現回数を計数する。

【０１１３】次に、ステップ１４０４において、出現回数取得プログラム１４６を起動し、上記特徴文字列抽出プログラム１４１で取得した特徴文字列のテキスト１５０中の各文書における出現回数を取得する。

【０１１４】最後に、ステップ１４０５において、類似度算出プログラム１４８を起動し、上記特徴文字列抽出プログラム１４１で取得した各特徴文字列に対する、上記種文書内出現回数取得プログラム１４５で取得した種文書内出現回数と、上記出現回数取得プログラム１４６で取得したテキスト１５０中の各文書における出現回数から、種文書とテキスト１５０内の各文書との類似度を算出する。

【０１１５】本実施例では、類似度の算出に、前述の類似度算出式（１）を用いるが、他の方法を用いても構わない。この類似度算出式（１）を用いて、前述の文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」が種文書として指定された場合の文書 iの類似度S(i)を算出すると、次のようになる。

【０１１６】S(1) = 1.0

S(2) = 0.262

S(3) = 0.048

S(4) = 0.0

以上が、類似文書検索プログラム１３１の処理手順である。

【０１１７】次に、図１４に示したステップ１４０２において、類似文書検索プログラム１３１により起動される特徴文字列抽出プログラム１４１の処理手順について、図１７のＰＡＤ図を用いて説明する。

【０１１８】特徴文字列抽出プログラム１４１では、ステップ１７００において、図１４に示したステップ１４０１における単一文字種文字列抽出プログラム１６１により、ワークエリア１７０に格納されている全ての単一文字種文字列を取得する。

【０１１９】次に、ステップ１７０１において、上記ステップ１７００で取得した全ての単一文字種文字列に対して、次のステップ１７０２～１７０４を繰り返し実行する。

【０１２０】すなわち、ステップ１７０２では、ステップ１７００で取得した単一文字種文字列の文字種を判定し、その文字種が漢字やカタカナである場合には、ステップ１７０３を実行し、漢字やカタカナ以外の場合には、ステップ１７０４を実行する。

【０１２１】ステップ１７０３では、後述する分割確率比較特徴文字列抽出プログラム１４２を起動し、漢字やカタカナの単一文字種文字列から特徴文字列を抽出する。

【０１２２】ステップ１７０４では、漢字やカタカナ以外の単一文字種文字列そのものを特徴文字列として抽出する。

【０１２３】そして、最後にステップ１７０５において、上記ステップ１７０２やステップ１７０３で抽出された特徴文字列をワークエリア１７０へ格納する。

【０１２４】以上が、特徴文字列抽出プログラム１４１の処理手順である。

【０１２５】以下、図１４に示した特徴文字列抽出プログラム１４１の処理手順について具体例を用いて説明する。

【０１２６】図２７に、前述の文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」から特徴文字列を抽出する例を示す。

【０１２７】まず、文書１から単一文字種文字列「・・・」「。」「携帯電話」「の」「使用時」「の」「マナー」「が」「問題」「になる」「。」「・・・」を抽出する。

【０１２８】次に、これらの単一文字種文字列の文字種を判定し、漢字文字列「携帯電話」、「使用時」および「問題」とカタカナ文字列「マナー」に対して分割確率比較特徴文字列抽出プログラム１４２により特徴文字列を抽出し、漢字文字列とカタカナ文字列以外の文字列「の」「の」「が」「になる」「。」からは単一文字種文字列そのものを特徴文字列として抽出する。

【０１２９】以上が、特徴文字列抽出プログラム１４１の具体的な処理例である。

【０１３０】次に、図１４に示したステップ１４０４において類似文書検索プログラム１３１により起動される出現回数取得プログラム１４６の処理手順を図２６のＰＡＤ図を用いて説明する。

【０１３１】出現回数取得プログラム１４６では、図１４に示したステップ１４０２においてワークエリア１７０に格納した特徴文字列を取得する（ステップ２６００）。

【０１３２】そして、ワークエリア１７０に格納されている全ての特徴文字列に対して、ステップ２６０２を実行する（ステップ２６０１）。

【０１３３】ステップ２６０２では、出現回数ファイル読込みプログラム１４７を起動し、テキスト１５０内の各文書における特徴文字列の出現回数を取得し、ワークエリア１７０に格納する。

【０１３４】以上が、出現回数取得プログラム１４６の処理手順である。

【０１３５】次に、図１７に示したステップ１７０３において特徴文字列抽出プログラム１４１により起動される分割確率比較特徴文字列抽出プログラム１４２の処理手順について、図１８のＰＡＤ図を用いて説明する。

【０１３６】分割確率比較特徴文字列抽出プログラム１４２は、ステップ１８００において、最後に特徴文字列が抽出された末尾の文字位置（以下、最新分割点と呼ぶ）LSの初期値を 0に設定する。

【０１３７】そして、図１７に示したステップ１７０３において、入力された単一文字種文字列の文字列長が予め定められた長さ以上のとき、次のステップ１８０２～１８０９までを繰り返し実行する（ステップ１８０１）。

【０１３８】ステップ１８０２では、後述する分割確率算出プログラム１４３を起動し、単一文字種文字列の先頭から i文字目の分割確率P(i)と、(i+1)文字目の分割確率P(i+1)を算出する。

【０１３９】次に、ステップ１８０３において、上記分割確率算出プログラム１４３で算出したP(i)とP(i+1)の値を比較し、P(i)がP(i+1)よりも大きい場合にはステップ１８０４を実行し、P(i)がP(i+1)よりも小さい場合にはステップ１８０６を実行し、P(i)とP(i+1)が等しい場合にはステップ１８０８を実行する。

【０１４０】ステップ１８０４では、単一文字種文字列の先頭から i文字目までの文字列を特徴文字列として抽出する。そして、ステップ１８０５において、最新分割点LSを iに設定し、 iの値を 1加算する。

【０１４１】ステップ１８０６では、単一文字種文字列の先頭から(i+1)文字目までの文字列を特徴文字列として抽出する。そして、ステップ１８０７において、最新分割点LSを(i+1)に設定し、iの値を 2加算する。

【０１４２】ステップ１８０８では、それぞれ単一文字種文字列の先頭から i文字目までの文字列と(i+1)文字目までの文字列を特徴文字列として抽出する。そして、ステップ１８０９において、最新分割点LSを(i+1)に設定し、 iの値を 2加算する。

【０１４３】以上が、分割確率比較特徴文字列抽出プログラム１４２の処理手順である。

【０１４４】以下、図１８に示した分割確率比較特徴文字列抽出プログラム１４２の処理手順について具体例を用いて説明する。

【０１４５】図8に、前述の文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」から抽出された単一文字種文字列「携帯電話」から特徴文字列を抽出する例を示す。

【０１４６】まず、「携帯電話」における1文字目での分割確率P(1)は、「携」の単独確率0.01と「帯」の先頭確率0.11の積として0.001が算出され、2文字目での分割確率P(2)は、「携帯」の単独確率0.10と「電」の先頭確率0.36の積として0.036が算出される。次に、これらの分割確率を比較し、値の大きい方で単一文字種文字列「携帯電話」を分割する。この場合、1文字目の分割確率P(1)(=0.000)よりも2文字目の分割確率P(2)(=0.036)の方が大きいので、単一文字種文字列「携帯電話」は「携帯」と「電話」に分割される。

【０１４７】また、図２０に、上記文書１から抽出した単一文字種文字列「マナー」から特徴文字列を抽出する例を示す。まず、「マナー」における2文字目での分割確率P(2)は、「マナ」の単独確率0.00と「ー」の単独確率0.00の積として0.00と算出される。次に、3文字目での分割確率P(3)、すなわち「マナー」が単一文字種文字列そのものとして出現する確率は「ナー」の末尾確率0.79と1.0の積として0.79と算出される。これらの値を比較して、値の大きい方で単一文字種文字列に分割される。この場合、「マナー」の2文字目での分割確率P(2)(=0.00)よりも3文字目での分割確率P(3)(=0.79)の方が大きいので、3文字目で分割されることになり、結果的に単一文字種文字列「マナー」は分割されないことになる。

【０１４８】以上が、分割確率比較特徴文字列抽出プログラム１４２の具体的な処理手順である。

【０１４９】次に、図１８に示したステップ１８０１において分割確率比較特徴文字列抽出プログラム１４２により起動される分割確率算出プログラム１４３の処理手順について、図１９のＰＡＤ図を用いて説明する。

【０１５０】分割確率算出プログラム１４３は、ステップ１９００において、図１８に示したステップ１８０１において指定される分割確率の算出位置 iおよび最新分割点LSを取得する。

【０１５１】次に、算出位置 iにおける分割確率P(i)を算出するために、ステップ１９０１～１９０６を実行し、各出現確率を取得する。

【０１５２】まず、ステップ１９０１において、図１２に示したステップ１２０１で抽出されたn-gramの長さ nと分割確率の算出位置 iを比較し、(i － LS)が n以下である場合には、ステップ１９０２を実行し、(i － LS)が nよりも大きい場合には、ステップ１９０３を実行する。

【０１５３】ステップ１９０２では、出現確率ファイル読込みプログラム１４４を起動し、最新分割点LSから i文字目までの文字列の単独確率を取得し、分割確率算出位置iの前方の文字列の出現確率Pre(i)とする。

【０１５４】ステップ１９０３では、出現確率ファイル読込みプログラム１４４を起動し、最新分割点LSから i文字目までの文字列の後方のn-gramの末尾確率を取得し、分割確率算出位置 iの前方の文字列の出現確率Pre(i)とする。

【０１５５】次に、ステップ１９０４において、単一文字種文字列の文字列長Lnと分割確率算出位置 iを比較し、Lnが(i+1)よりも大きい場合にはステップ１９０５を実行し、Lnが(i+1)と等しい場合には、ステップ１９０６を実行する。

【０１５６】ステップ１９０５では、出現確率ファイル読込みプラグラム１４４を起動し、(i+1)文字目のtgram

の先頭確率を取得し、分割確率算出位置 iの後方の文字列の出現確率Post(i)とする。

【０１５７】ステップ１９０６では、出現確率ファイル読込みプラグラム１４４を起動し、(i+1)文字目の1gramの単独確率を取得し、分割確率算出位置 iの後方の文字列の出現確率Post(i)とする。

【０１５８】次に、算出位置(i+1)における分割確率P(i+1)を算出するために、ステップ１９０７～１９１３を実行し、各出現確率を取得する。

【０１５９】まず、ステップ１９０７において、図１２に示したステップ１２０１で抽出されたn-gramの長さ nと分割確率の算出位置 iを比較し、((i+1) - LS)が n以下である場合には、ステップ１９０８を実行し、((i+1) - LS)が nよりも大きい場合には、ステップ１９０９を実行する。

【０１６０】ステップ１９０８では、出現確率ファイル読込みプログラム１４４を起動し、最新分割点LSから(i+1)文字目までの文字列の単独確率を取得し、分割確率算出位置(i+1)の前方の文字列の出現確率Pre(i+1)とする。

【０１６１】ステップ１９０９では、出現確率ファイル読込みプログラム１４４を起動し、最新分割点LSから(i+1)文字目までの文字列の後方のn-gramの末尾確率を取得し、分割確率算出位置(i+1)の後方の文字列の出現確率Pre(i+1)とする。

【０１６２】次に、ステップ１９１０において、単一文字種文字列の文字列長Lnと分割確率算出位置 iを比較し、Lnが(i+2)よりも大きい場合にはステップ１９１１を実行し、Lnが(i+2)と等しい場合には、ステップ１９１２を実行し、Lnが(i+1)と等しい場合には、ステップ１９１３を実行する。

【０１６３】ステップ１９１１では、出現確率ファイル読込みプラグラム１４４を起動し、(i+2)文字目の1gramの先頭確率を取得し、分割確率算出位置(i+1)の後方の文字列の出現確率Post(i+1)とする。

【０１６４】ステップ１９１２では、出現確率ファイル読込みプラグラム１４４を起動し、(i+2)文字目の1gramの単独確率を取得し、分割確率算出位置(i+1)の後方の文字列の出現確率Post(i+1)とする。

【０１６５】ステップ１９１３では、分割確率算出位置(i+1)の後方の文字列の出現確率Post(i+1) = 1とする。

【０１６６】次に、ステップ１９１４において、上記ステップ１９０１～１９０３で取得したPre(i)と上記ステップ１９０４～１９０６で取得したPost(i)の積を算出位置iにおける分割確率P(i)とし、上記ステップ１９０７～１９０９で取得したPre(i+1)と上記ステップ１９１０～１９１３で取得したPost(i+1)の積を算出位置(i+1)における分割確率P(i+1)として、それぞれワークエリア１７０に格納する。

【０１６７】以上が、分割確率算出プログラム１４３の処理手順である。

【０１６８】以下、図１９に示した分割確率算出プログラム１４３の処理手順について具体例を用いて説明する。

【０１６９】図２８に前述の文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」から抽出された単一文字種文字列「携帯電話」の分割確率を算出する例を示す。なお、本図に示す例では、出現確率ファイル１５２に格納されているn-gram長を 2とし、分割確率を算出する i文字目を 1文字目とする。すなわち、 1文字目での分割確率P(1)および 2文字目での分割確率P(2)を算出するものとして、以下の説明を行なう。

【０１７０】まず、分割確率の算出位置である 1文字目までの文字列の単独確率が出現確率ファイル６００に格納されているかどうかを確認するために、出現確率ファイル６００に格納されているn-gram長 2と分割確率算出位置 1を比較する。その結果、格納されているn-gram長の方が大きいので、 1文字目までの文字列「携」の単独確率0.01を出現確率ファイル６００より取得する。

【０１７１】次に、分割確率の算出位置の後方に何文字存在するかを確認するために、単一文字種文字列「携帯電話」の文字列長 4と分割確率算出位置 1を比較する。その結果、 2文字以上の文字列「帯電話」が存在するため、「帯」の先頭確率0.11を出現確率ファイル６００から取得する。そして、「携」の単独確率0.01と「帯」の先頭確率0.11の積を算出し、 1文字目での分割確率P(1)=0.001を得る。

【０１７２】同様に、分割確率の算出位置である 2文字目までの文字列の単独確率が出現確率ファイル６００に格納されているかを確認するために、出現確率ファイル６００に格納されているn-gram長 2と分割確率算出位置 2を比較する。その結果、格納されているn-gram長と算出位置が等しいので、 2文字目までの文字列「携帯」の単独確率0.10を出現確率ファイル６００より取得する。

【０１７３】次に、分割確率の算出位置の後方に何文字存在するかを確認するために、単一文字種文字列「携帯電話」の文字列長 4と分割確率算出位置 2を比較する。その結果、 2文字以上の文字列「電話」が存在するため、「電」の先頭確率0.36を出現確率ファイル６００から取得する。そして、「携帯」の単独確率0.10と「帯」の先頭確率0.36の積を算出し、2文字目での分割確率P(2)=0.036を得る。

【０１７４】以上が、分割確率算出プログラム１４３の具体的な処理手順である。

【０１７５】以上が、本発明の第一の実施例である。

【０１７６】本実施例では、出現情報ファイル１５１と出現確率ファイル１５２に格納するn-gramの長さとして２を用いて、特徴文字列抽出プログラム１４３の処理手順を説明したが、この長さとして１や３等の固定値を用いてもよいし、データベース中の出現回数等の情報に基

づき可変長としてもよいし、単一文字種文字列自体の長さｍとしてもよいし、さらには、それらの組み合わせであっても、同様に特徴文字列抽出の処理を行なうことができるのは明らかであろう。

【０１７７】また、本実施例では、種文書の内容に類似する文書を検索するものとして特徴文字列抽出プログラム１４３の処理手順を説明したが、この種文書の代わりに、文章が指定されたとしても同様に特徴文字列を抽出することができ、類似文書検索を行なうことができるのは明らかであろう。

【０１７８】また、本実施例では、単一文字種文字列の先頭から n文字目までの分割確率と(n+1)文字目までの分割確率を比較することで特徴文字列を抽出する例を用いて、分割確率比較特徴文字列抽出プログラム１４２の処理手順を説明したが、単一文字種文字列の末尾から、それぞれ n文字目までの分割確率と(n+1)文字目までの分割確率を比較しても、さらには、単一文字種文字列中の m文字(mは 1以上の整数)と n文字の分割確率を比較しても、同様に、文書の特徴を表す特徴文字列の抽出が行えることは明らかであろう。

【０１７９】なお、本実施例においては、漢字やカタカナの単一文字種文字列に対する分割確率比較特徴文字列抽出プログラム１４２を含む構成として説明したが、漢字あるいはカタカナを含まないデータベースを対象とする場合等には、対応する分割確率比較特徴文字列抽出プログラム１４２を含まない構成としてもよいし、漢字やカタカナ以外に対応する分割確率比較特徴文字列抽出プログラム１４２を含む構成としてもよいし、従来技術２で示したように、各文字種に対応する特徴文字列抽出プログラムを含む構成であってもよい。

【０１８０】また、本実施例においては、単一文字種文字列から特徴文字列を抽出する構成としたが、特定の文字種間を境界として前後に跨る部分文字列から特徴文字列を抽出することにより、例えば、「Ｆ１」や「ビタミンＣ」、「W杯」、「ケイ素」等の文字列を検索に用いることができ、さらに高精度な類似文書検索を実現することも可能となる。

【０１８１】また、本実施例における出現情報ファイル作成登録プログラム１２１では、文字種境界を単語の区切れ目とみなし、各n-gramの先頭回数、末尾回数および単独回数を計数するものとしたが、付属語、すなわち助詞や助動詞等を単語の区切れ目の候補とみなし、各n-gramの先頭回数、末尾回数および単独回数を計数してもよい。

【０１８２】さらに、本実施例においては、出現情報ファイル１５１を図５に示した表形式で作成されるものとしたが、この方法では、対象とするn-gram長が増大するにともない、n-gram種類が増加するため、分割確率ファイル作成登録プログラム１２４の処理に長大な時間を要することになる。この問題は、特徴文字列に対して、検索用のインデクスを付加することにより解決できる。これにより、n-gram種類が増加しても、高速に登録処理を実現することができる。この特徴文字列に対する検索用インデクスとしては、全文検索用インデクス１５３を用いてもよいし、「特開平８－３２９１１２号公報」等に開示されているような単語インデクス方式を用いてもよい。この問題は、出現確率ファイル１５２および出現回数ファイル１５３においても発生するが、同様に検索用のインデクスを付加することで解決することができる。

【０１８３】さらに、本実施例においては、文書登録時に出現確率ファイル作成登録プログラム１２４を起動し、出現確率ファイル１５２を作成する構成としたが、類似文書検索時の分割確率比較特徴文字列抽出プログラム１４２実行時に、出現情報ファイル１５１に格納されている各n-gramの出現情報から該当する出現確率を算出することにより、磁気ディスク１０５に格納するファイルを削減することも可能である。

【０１８４】また、本実施例においては、特徴文字列抽出プログラム１４１により抽出された特徴文字列を用いた類似文書検索システムについて説明したが、種文書から特徴文字列を抽出する特徴文字列抽出システムとして用いることも可能であるし、「特開平８－１５３１２１号公報」に示されるような形態素解析により文書中に含まれる単語を抽出し、これを用いて文書を自動的に分類するシステムに用いることも可能である。

【０１８５】ただし、第一の実施例における分割確率比較特徴文字列抽出プログラム１４２は、 i文字目での分割確率P(i)と(i+1)文字目での分割確率P(i+1)を比較し、その値の大きい方で分割するため、全ての単一文字種文字列から(i+1)文字以下の特徴文字列が抽出されてしまい、(i+1)文字より長い単語が誤って分割されてしまうという問題がある。

【０１８６】以下、第一の実施例で(i+1)文字より長い単語が誤って分割されてしまうという問題が生じる例を図２２に示す具体例を用いて説明する。なお、本図では、漢字で構成される単一文字種文字列「北海道」を対象とし、分割確率算出位置 iの初期値を 1とする。

【０１８７】分割確率比較特徴文字列抽出プログラム１４２では、まず、ステップ２２００において、前述した分割確率算出プログラム１４３を起動し、1文字目の分割確率P(1)と2文字目の分割確率P(2)を算出する。本図に示した例では、単一文字種文字列「北海道」の1文字目で「北」と「海道」に分割される確率は、1-gram「北」の単独確率0.03と2-gram「海道」の単独確率0.00の積としてP(1)=0.000と算出される。同様に、2文字目で「北海」と「道」に分割される確率は、2-gram「北海」の単独確率0.03と1-gram「道」の単独確率0.12の積P(2)=0.004として算出される。

【０１８８】次に、ステップ２２０１において、上記ステップ２２００で算出されたP(1)とP(2)のうち、値の大

きい方を分割点とし、単一文字種文字列の先頭から分割点までの文字列を特徴文字列として抽出する。本図に示した例では、P(2)の方がP(1)よりも大きいので、 2文字目で単一文字種文字列「北海道」を分割し、 2文字目までの文字列“北海”を特徴文字列として抽出する。

【０１８９】次に、ステップ２２０２において、最後に特徴文字列が抽出された末尾の文字位置（以下、最新分割点と呼ぶ）LSを 2に設定し、最新分割点以降の単一文字種文字列「道」を対象に特徴文字列抽出処理を継続する。

【０１９０】次に、ステップ２２０３において、単一文字種文字列「道」の文字列長 1は、予め定められた長さ 2未満であるため、文字列“道”が特徴文字列として抽出される。この結果、「・・・。道の駅と呼ばれるサービスエリアが国道沿いに建設されることになった。・・・」等という文書が類似文書として誤って検索されてしまうことになる。

【０１９１】以上が、第一の実施例における分割確率比較特徴文字列抽出プログラム１４２の処理例である。本図に示した例では、 1文字目と 2文字目の分割確率P(1)とP(2)を比較し、値の大きい方を分割点とするため、単一文字種文字列「北海道」から“北海”と“道”が特徴文字列として抽出されてしまい、種文書の中心概念からずれた文書が類似文書として検索されてしまう。

【０１９２】このために、本発明を適用した類似文書検索システムの第二の実施例では、単一文字種文字列から特徴文字列を抽出する際に算出された分割確率が所定値（以下、分割閾値と呼ぶ）よりも高い場合にのみ、比較処理を行なうことにより、(i+1)文字より長い特徴文字列を抽出できるようにする。

【０１９３】本実施例は、第一の実施例（図１）とほぼ同様の構成を取るが、分割確率比較特徴文字列抽出プログラム１４２の処理手順が異なり、図２１のＰＡＤ図に示すように、ステップ２１００～２１０４が追加される。

【０１９４】以下、第二の実施例における分割確率比較特徴文字列抽出プログラム１４２ａの処理手順について図２１のＰＡＤ図を用いて説明する。

【０１９５】分割確率比較特徴文字列抽出プログラム１４２ａでは、ステップ１８００において、最新分割点LSの初期値を 0に設定する。

【０１９６】そして、特徴文字列の抽出対象となる単一文字種文字列の文字列長が予め定められた長さ以上のとき、次のステップ１８０２～１８０７、ステップ２１０１～２１０３までを繰り返し実行する（ステップ２１００）。

【０１９７】ステップ１８０２では、分割確率算出プログラム１４３を起動し、単一文字種文字列の先頭から i文字目の分割確率P(i)と、(i+1)文字目の分割確率P(i+1)を算出する。

【０１９８】次に、ステップ２１００において、上記分割確率算出プログラム１４３で算出された分割確率P(i)、P(i+1)の値および予め定められた分割閾値Thの値を比較し、最大のものを抽出する。この結果、分割確率P(i)が抽出されたならばステップ１８０４を実行し、分割確率P(i+1)が抽出された場合にはステップ１８０６を実行し、分割閾値Thが抽出された場合にはステップ２１０１を実行する。

【０１９９】ステップ１８０４では、単一文字種文字列の先頭から i文字目までの文字列を特徴文字列として抽出する。そして、ステップ１８０５において、最新分割点LSを iに設定し、 iの値を 1加算する。

【０２００】ステップ１８０６では、単一文字種文字列の先頭から(i+1)文字目までの文字列を特徴文字列として抽出する。そして、ステップ１８０７において、最新分割点LSを(i+1)に設定し、iの値を 2加算する。

【０２０１】ステップ２１０１では、分割確率の算出位置 iと単一文字種文字列の文字列長Lnとを比較し、(i+1)が文字列長Lnよりも小さい場合には、 ステップ２１０２を実行し、(i+1)が文字列長Ln以上であるならば、ステップ２１０３を実行する。

【０２０２】ステップ２１０２では、分割確率の算出位置 iの値を 1加算する。

【０２０３】ステップ２１０３では、単一文字種文字列そのものを特徴文字列として抽出する。そして、ステップ２１０４において、最新分割点LSを文字列長Lnに設定し、iの値を 1加算する。

【０２０４】以上が、分割確率比較特徴文字列抽出プログラム１４２ａの処理手順である。

【０２０５】以下、第二の実施例における分割確率比較特徴文字列抽出プログラム１４２ａの処理手順をそれぞれ図２３に示す具体例で説明する。なお、本図では、漢字で構成される単一文字種文字列「北海道」を対象とし、分割閾値Thを0.050とし、分割確率算出位置 iの初期値を 1として分割確率比較特徴文字列抽出プログラム１４２ａの処理手順を説明する。

【０２０６】分割確率比較特徴文字列抽出プログラム１４２ａでは、まず、ステップ２２００において、前述した分割確率算出プログラム１４３を起動し、 1文字目の分割確率P(1)と 2文字目の分割確率P(2)を算出し、P(1)=0.000およびP(2)=0.004を得る。

【０２０７】次にステップ２３０１において、上記ステップ２２００で算出した分割確率P(1)、P(2)および分割閾値Thうち、最大のものを抽出する。この結果、分割閾値Thが最大であるので、ステップ２３０２において、分割確率の算出位置 i(=1)と単一文字種文字列「北海道」の文字列長Ln(=3)を比較する。この結果、分割確率の算出位置 iの方が小さいので、 iの値を 1加算する。

【０２０８】そして、ステップ２３０４において、2文字目での分割確率P(2)と3文字目での分割確率P(3)を算

出する。この例では、2文字目で「北海」と「道」に分割される確率は、2-gram「北海」の単独確率0.03と1-gram「道」の単独確率0.12の積P(2)=0.004として算出され、3文字目までの「北海道」として出現する確率は、2-gram「北海」の先頭確率と2-gram「海道」の末尾確率の積P(3)=0.465として算出される。

【０２０９】次に、ステップ２３０５において、上記ステップ２３０４で算出した分割確率P(2)、P(3)および分割閾値Thのうち、最大のものを抽出する。この結果、P(3)が最大であるので、3文字目「北海道」までが特徴文字列として抽出される。

【０２１０】以上説明したように、本実施例によれば、分割確率が分割閾値よりも高い場合にのみ、比較処理を行なうようにすることにより、本来分割されることのない位置での分割を削減することができる。このため、第一の実施例で抽出されていた不適切な特徴文字列を大幅に削減することが可能となる。そのため、種文書が表わす概念を適切に検索できるとともに、高速に類似文書を検索することができるようになる。

【０２１１】次に、本発明の第三の実施例について図２９を用いて説明する。

【０２１２】第一の実施例および第二の実施例においては、特徴文字列として抽出される可能性のある全ての文字列を出現回数ファイル１５３中に格納しておく必要があるため、文字列の種類の増加に伴い、データベース内の各文書における出現回数の取得に長大な時間を要するとともに、必要な磁気ディスク容量が増加してしまう。

【０２１３】本発明を適用した類似文書検索システムの第三の実施例は、種文書から抽出した特徴文字列に対するデータベース内の各文書における出現回数の取得に、出現回数ファイル１５３を用いずに、全文検索用インデクスを利用することにより上記必要な磁気ディスク容量を低減する方式である。

【０２１４】すなわち、本実施例によれば、第一の実施例におけるデータベース内の各文書における出現回数の取得に全文検索システムを利用することにより、文字列の種類数が多いデータベースに対しても高速な類似文書検索を実現することが可能となる。さらに、出現回数ファイル１５３を全文検索用インデクスで代用するため、本類似文書検索システムを全文検索システムと組み合わせて実現した場合に、第一の実施例に比べ必要となる磁気ディスク容量を削減できることになる。

【０２１５】本実施例は、第一の実施例（図１）とほぼ同様の構成を取るが、類似文書検索プログラム１３１中の出現回数取得プログラム１４６を構成する出現回数ファイル読込みプログラム１４７が異なる。このプログラムの代わりに、図２９に示すように全文検索プログラム２９０２が用いられる。

【０２１６】以下、本実施例における処理手順のうち、第一の実施例とは異なる出現回数取得プログラム１４６ａの処理手順について、図３０を用いて説明する。

【０２１７】ここで、第一の実施例における出現回数取得プログラム１４６（図２６）と異なる点は、出現回数取得ステップ３０００だけである。他の処理ステップの処理手順は、第一の実施例で説明した通りである。

【０２１８】出現回数取得ステップ３０００では、特徴文字列抽出プログラム１４１によりワークエリア１７０に格納された特徴文字列を全文検索プログラム２９０２で検索することにより、テキスト１５０内の各文書における該特徴文字列の出現回数を取得する。

【０２１９】本実施例の出現回数取得ステップ３０００で用いる全文検索プログラム２９０２としては、どのような方式を適用しても構わない。例えば、「特開昭６４－３５６２７号公報」（以下、従来技術４と呼ぶ）で開示されているようなn-gramインデクス方式を用いることも可能である。

【０２２０】この従来技術４によるn-gramインデクス方式では、図２９に示すように、文書の登録時に、データベースへ登録する文書のテキストデータからn-gramとそのn-gramのテキスト中における出現位置を抽出し、全文検索用インデクス２９０１として磁気ディスク装置２９００に格納しておく。検索時には指定された検索ターム中に出現するn-gramを抽出し、これらに対応するインデクスを上記磁気ディスク装置２９００中の全文検索用インデクス２９０１から読み込み、インデクス中のn-gramの出現位置を比較し、検索タームから抽出したn-gramの位置関係とインデクス中のn-gramの位置関係が等しいかどうかを判定することによって、指定された検索タームが出現する文書を高速に検索する。

【０２２１】この方法を用いて、特徴文字列を検索タームとして全文検索プログラム２９０２へ入力し、該特徴文字列の出現文書とその位置情報を取得することにより、該特徴文字列の各文書における出現回数を取得することが可能となる。

【０２２２】以下、この従来技術４を用いた出現回数の取得方法を図７と図１５を用いて具体的に説明する。なお、本図では、n-gramの nの値を 1としている。

【０２２３】まず、文書の登録時の処理手順を図７を用いて具体的に説明する。データベースに登録するテキスト７０１がn-gramインデクス作成登録ステップ７０２に読み込まれ、n-gramインデクス７００が作成される。このn-gramインデクス７００には、テキスト７０１に出現する全ての1-gramとテキスト７０１における1-gramの出現位置が格納される。

【０２２４】本図に示すテキスト７０１では、「携」という1-gramはテキスト７０１内の文書番号２の２６文字目に現れるので、n-gramインデクス７００には1-gram「携」とこれに対応したかたちで、出現位置（２，２６）が格納される。ここで、例えば、（２，２６）は、文書番号２の２６文字目に出現するということを示して

いる。

【０２２５】次に、検索時の処理手順を図１５を用いて具体的に説明する。本図では、前述の文書１「携帯電話の使用のマナーが問題になる。・・・」から抽出された特徴文字列「電話」の出現回数を、前述したn-gramインデクス７００から取得する例について示す。

【０２２６】まず、検索対象となる特徴文字列がn-gram抽出部１５００に入力され、特徴文字列中に出現する全ての n-gramとその n-gramの特徴文字列における出現位置が抽出される。次に、抽出された n-gramとこれに対応する n-gramの特徴文字列における出現位置がインデクス検索部１５０１に入力される。インデクス検索部１５０１では、特徴文字列から抽出された n-gramに対応するインデクスが n-gramインデクス７００から読み込まれ、これらのインデクスの中から文書番号が一致し、かつ特徴文字列中の位置関係と同じ位置関係を持つものが抽出され、検索結果として出力される。

【０２２７】特徴文字列として「電話」が入力された本図の場合、まず、n-gram抽出部１５００において、（1-gram「電」、1-gram位置「１」）と（1-gram「話」、1-gram位置「２」）が抽出される。ここで、 n-gram位置「１」は検索タームの先頭、n-gram位置「２」はその次の文字位置を示す。

【０２２８】次に、インデクス検索部１５０１において、n-gramインデクス７００から1-gram「電」と「話」に対応するインデクスが読み込まれる。これらのインデクスにおける出現文書番号が等しく、かつ出現位置がn-gram位置「１」とn-gram位置「２」のように連続するものが、すなわち隣接するものが抽出され検索結果として出力される。

【０２２９】本図では、 1-gram「電」の（２，２８）と1-gram「話」の出現位置（２，２９）が文書番号が同じで、位置が「２８」と「２９」で隣接するため、 n-gram「電話」が文字列として存在することが分かり、文書２中に検索ターム「電話」が出現することが検出される。しかし、 1-gram「電」の（３，１１）と1-gram「話」の（３，１５）は隣接していないため、この位置には特徴文字列「電話」が出現しないことが分かる。

【０２３０】そして、上記インデクス検索部１５０１から検索結果として出力される出現位置を計数することにより、該当特徴文字列の出現回数を得る。

【０２３１】以上説明したように、本実施例によれば、出現回数ファイルの特徴文字列検索用インデクスと出現回数ファイルの代わりに、全文検索用インデクスを利用することにより、余分なファイルを増やさずに、高速に類似文書検索を実現することが可能となる。

【０２３２】次に、本発明の第四の実施例について図３１を用いて説明する。

【０２３３】第一、第二および第三の実施例においては、種文書から抽出された単一文字種文字列の先頭からn文字目での分割確率と(n+1)文字目での分割確率を比較することで特徴文字列を抽出するものとしたが、出現情報ファイル１５１と出現確率ファイル１５２を保持する必要があるため、文字列の種類の増加に伴い、必要な磁気ディスク容量が増加してしまう。

【０２３４】本発明を適用した類似文書検索システムの第四の実施例は、出現情報ファイル１５１と出現確率ファイル１５２を用いずに、出現回数ファイル１５３を利用することで、上記必要な磁気ディスク容量を低減する方式である。

【０２３５】本発明を適用した第四の実施例は、第一の実施例（図１）とほぼ同様の構成をとるが、類似文書検索プログラム１３１を構成する特徴文字列抽出プログラム１４１が異なり、n-gram抽出プログラム３１００と前述の出現回数取得プログラム１４６で構成される。

【０２３６】以下、本実施例における処理手順のうち、第一の実施例とは異なる特徴文字列抽出プログラム１４１ａの処理手順について、図３２を用いて説明する。

【０２３７】特徴文字列抽出プログラム１４１ａは、まずステップ３２００において、前述の単一文字種文字列抽出プログラム１６１により、ワークエリア１７０に格納されている全ての単一文字種文字列を取得する。

【０２３８】次に、ステップ３２０１において、上記ステップ３２００で取得した全ての単一文字種文字列に対して、次のステップ３２０２～３２０５を繰り返し実行する。

【０２３９】すなわち、ステップ３２０２では、n-gram抽出プログラム３１００を起動し、ステップ３２００で取得した単一文字種文字列から、予め定められた長さ n（ nは 1以上の整数）のn-gramを先頭から 1文字ずつずらしながら、全てのn-gramを抽出する。

【０２４０】そして、ステップ３２０３において、上記n-gram抽出プログラム３１００により抽出された全てのn-gramに対して、次のステップ３２０４を繰り返し実行する。すなわち、ステップ３２０４では、出現回数取得プログラム１４６を起動し、上記n-gram抽出プログラム３１００により抽出されたn-gramの出現回数を取得する。

【０２４１】そして、ステップ３２０５において、上記ステップ３２０４で取得した各n-gramの出現回数の降順にソートし、上位から予め定められた個数のn-gramを特徴文字列として抽出する。

【０２４２】以上が、特徴文字列抽出プログラム１４１ａの処理手順である。

【０２４３】以下、図３２に示した特徴文字列抽出プログラム１４１ａの処理手順について具体例を用いて説明する。

【０２４４】図３３に、前述の文書１「・・・。携帯電話の使用時のマナーが問題になる。・・・」から特徴文字列を抽出する例を示す。本図に示す例ではn-gramのn

の値を2とし、各単一文字種文字列から2個の2-gramを特徴n-gramとして抽出するものとする。

【0245】まず、文書1から単一文字種文字列「・・・」「。」「携帯電話」「の」「使用時」「の」「マナー」「が」「問題」「になる」「。」「・・・」を抽出する。

【0246】次に、これらの単一文字種文字列の先頭から1文字ずつずらしながら全ての2-gramを抽出し、各2-gramの出現回数の降順にソートする。例えば、単一文字種文字列「携帯電話」からは“携帯”、“帯電”、“電話”の3つの2-gramを抽出し、それぞれデータベース内の出現回数を取得する。この結果、（電話，5,283）、（携帯，462）、（帯電，269）が得られる。ここで（電話，5,282）は、2-gram“電話”のデータベース内における出現回数が5,283回であることを表わす。

【0247】次に、各単一文字種文字列において、上位2個の2-gramを特徴n-gramとして抽出する。この結果、単一文字種文字列「携帯電話」では（電話，5,283）、（携帯，462）が上位2個であるため、“電話”および“携帯”が特徴文字列として抽出される。

【0248】以上が、特徴文字列抽出プログラム141aの具体的な処理例である。

【0249】以上説明したように、本実施例によれば、出現情報ファイル151と出現確率ファイル152を用いずに、出現回数ファイル153を利用することにより、データベース中での実際の出現状況を正確に反映した特徴文字列を抽出することが可能となる。

【0250】なお、本実施例では、単一文字種文字列の先頭から 1文字ずつずらしながら予め定められた長さ nのn-gramを全て抽出するものとして、n-gram抽出プログラム3100の処理手順を説明したが、単一文字種文字列中の任意のn-gramを抽出してもよいし、さらには、単一文字種文字列中の m-gram(mは 1以上の整数)と n-gramを抽出してもよい。さらに、抽出するn-gramの長さ nを予め定められたものとしたが、単一文字種文字列の長さにより抽出する nの値を変更してもよいし、単一文字種文字列の文字種により変更してもよい。また、本発明のn-gram抽出手法は、文書の特徴を表すn-gramを抽出することができるため、n-gramを用いた文書の特徴を表すベクトルの算出やn-gramを用いた文書の分類にも適用可能である。

【0251】

【発明の効果】本発明によれば、誤分割が少なくなるように特徴文字列を抽出することができるようになる。これにより、単語辞書を参照せずに類似文書検索を行なった場合でも、意味のまとまった文字列を用いて検索を行なうことができるため、中心概念のずれを低減した類似文書検索を実現できる。

【図面の簡単な説明】

【図1】本発明による類似文書検索システムの第一の実施例の全体構成を示す図である。

【図2】従来技術3における出現情報ファイルの例を示す図である。

【図3】従来技術3における出現確率ファイルの例を示す図である。

【図4】従来技術3における特徴文字列抽出方法の例を示す図である。

【図5】本発明による出現情報ファイルの例を示す図である。

【図6】本発明による出現確率ファイルの例を示す図である。

【図7】本発明の第三の実施例におけるn-gramインデクスの例を示す図である。

【図8】本発明の第一の実施例における分割確率比較特徴文字列抽出プログラム142を漢字文字列に適用した場合の処理例を示す図である。

【図9】本発明による特徴文字列の抽出方法の例を示す図である。

【図10】本発明の第一の実施例におけるシステム制御プログラム110の処理手順を示すPAD図である。

【図11】本発明の第一の実施例における文書登録制御プログラム111の処理手順を示すPAD図である。

【図12】本発明の第一の実施例における出現情報ファイル作成登録プログラム121の処理手順を示すPAD図である。

【図13】本発明の第一の実施例における検索制御プログラム112の処理手順を示すPAD図である。

【図14】本発明の第一の実施例における類似文書検索プログラム131の処理手順を示すPAD図である。

【図15】本発明の第三の実施例における出現回数取得の例を示す図である。

【図16】本発明の第一の実施例における出現確率ファイル作成登録プログラム124の処理手順を示すPAD図である。

【図17】本発明の第一の実施例における特徴文字列抽出プログラム141の処理手順を示すPAD図である。

【図18】本発明の第一の実施例における分割確率比較特徴文字列抽出プログラム142の処理手順を示すPAD図である。

【図19】本発明の第一の実施例における分割確率算出プログラム143の処理手順を示すPAD図である。

【図20】本発明の第一の実施例における分割確率比較特徴文字列抽出プログラム142をカタカナ文字列に適用した場合の処理例を示す図である。

【図21】本発明の第二の実施例における分割確率比較特徴文字列抽出プログラム142aの処理手順を示すPAD図である。

【図22】本発明の第一の実施例における分割確率比較特徴文字列抽出プログラム142の処理例を示す図である。

【図２３】本発明の第二の実施例における分割確率比較特徴文字列抽出プログラム１４２ａの処理例を示す図である。
【図２４】本発明による出現回数ファイル作成処理の手順を示す図である。
【図２５】本発明の第一の実施例における出現回数ファイル作成登録プログラム１２７の処理手順を示すＰＡＤ図である。
【図２６】本発明の第一の実施例における出現回数取得プログラム１４６の処理手順を示すＰＡＤ図である。
【図２７】本発明の第一の実施例における特徴文字列抽出プログラム１４１の処理例を示す図である。
【図２８】本発明の第一の実施例における分割確率算出の処理例を示す図である。
【図２９】本発明の第三の実施例における類似文書検索プログラム１３１の構成を示す図である。
【図３０】本発明の第三の実施例における出現回数取得プログラム１４６ａの処理手順を示す図である。
【図３１】本発明の第四の実施例における特徴文字列抽出プログラム１４１ａの構成を示す図である。
【図３２】本発明の第四の実施例における特徴文字列抽出プログラム１４１ａの処理手順を示すＰＡＤ図である。
【図３３】本発明の第四の実施例における特徴文字列抽出プログラム１４１ａの処理例を示す図である。
【符号の説明】
１００…ディスプレイ、
１０１…キーボード、
１０２…中央演算処理装置（ＣＰＵ）、
１０３…フロッピディスクドライブ（ＦＤＤ）、
１０４…フロッピディスク、
１０５…磁気ディスク装置、
１０６…主メモリ、
１０７…バス、
１１０…システム制御プログラム、
１１１…文書登録制御プログラム、
１１２…検索制御プログラム、
１２０…テキスト登録プログラム、
１２１…出現情報ファイル作成登録プログラム、
１２２…出現情報計数プログラム、
１２３…出現情報ファイル作成プログラム、
１２４…出現確率ファイル作成登録プログラム、
１２５…出現確率算出プログラム、
１２６…出現確率ファイル作成プログラム、
１２７…出現回数ファイル作成登録プログラム、
１２８…出現回数計数プログラム、
１２９…出現回数ファイル作成プログラム、
１３０…検索条件式解析プログラム、
１３１…類似文書検索プログラム、
１３２…検索結果出力プログラム、
１４０…種文書読込みプログラム、
１４１…特徴文字列抽出プログラム、
１４２…分割確率比較特徴文字列抽出プログラム、
１４３…分割確率算出プログラム、
１４４…出現確率ファイル読込みプログラム、
１４５…種文書内出現回数計数プログラム、
１４６…出現回数取得プログラム、
１４７…出現回数ファイル読込みプログラム、
１４８…類似度算出プログラム、
１５０…テキスト、
１５１…出現情報ファイル、
１５２…出現確率ファイル、
１５３…出現回数ファイル、
１６０…共有ライブラリ、
１６１…同一文字種文字列抽出プログラム、
１７０…ワークエリア

【図２】

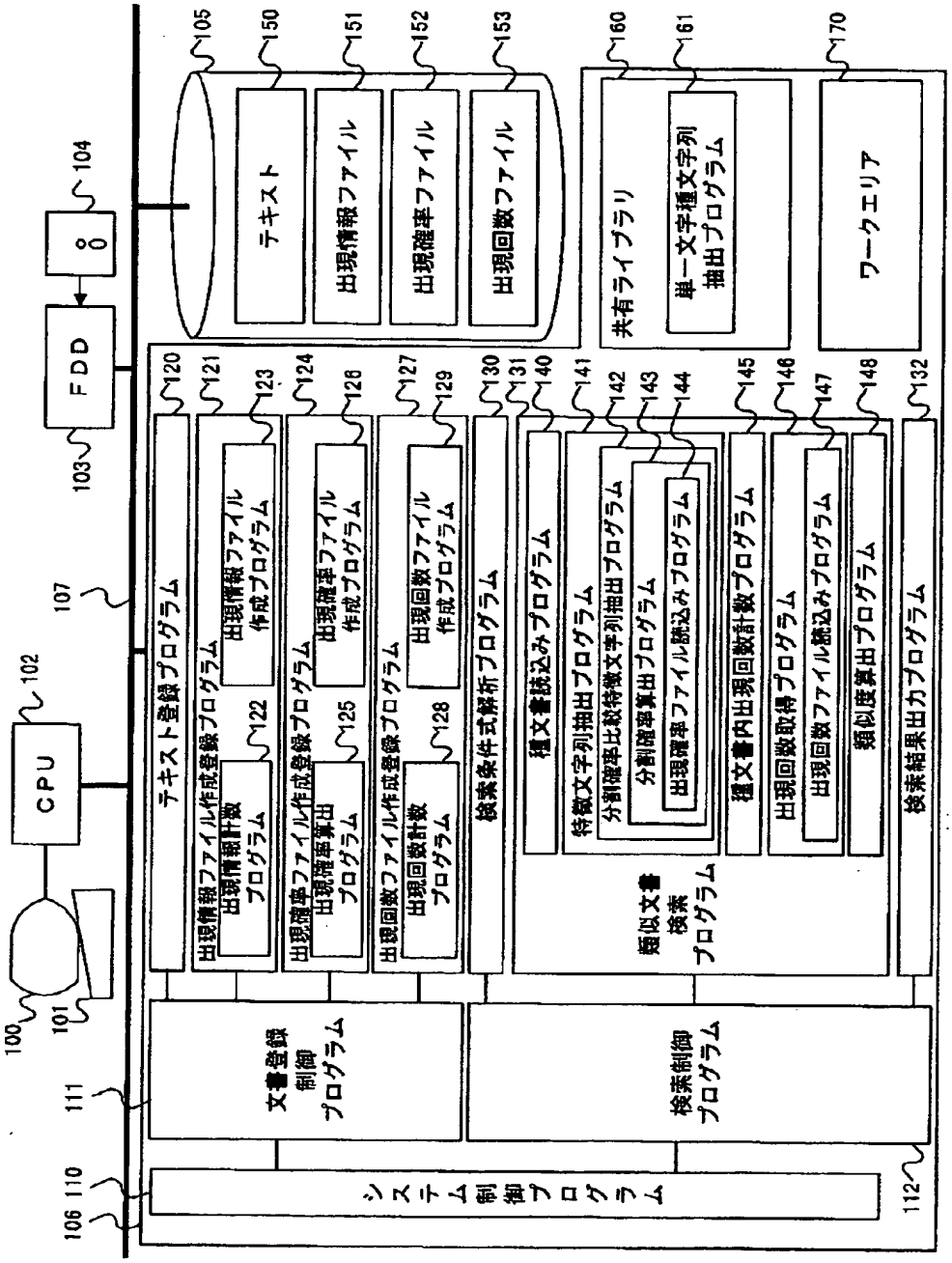
図２

No.	1-gram	出現回数	先頭回数	末尾回数
1	一	62,318	0	13,480
2	ナ	28,090	2,653	2,079
3	マ	43,300	15,235	6,179
4	携	4,740	768	492
5	帯	4,703	530	687
6	題	36,338	733	32,342
7	電	38,317	13,794	3,218
8	問	46,216	19,205	11,884
9	用	59,987	5,132	33,600
10	話	18,416	1,105	6,353

～ 200

【図1】

図1


CPU
FDD
テキスト登録プログラム
出現情報ファイル作成プログラム
出現情報計数プログラム
出現確率ファイル作成プログラム
出現確率算出プログラム
出現回数ファイル作成プログラム
出現回数計数プログラム
検索条件式解析プログラム
種文書読込みプログラム
特徴文字列抽出プログラム
分割確率比較特徴文字列抽出プログラム
分割確率算出プログラム
出現確率ファイル読込みプログラム
種文書内出現回数計数プログラム
出現回数取得プログラム
出現回数ファイル読込みプログラム
類似度算出プログラム
検索結果出力プログラム
類似文書検索プログラム
文書登録制御プログラム
検索制御プログラム
システム制御プログラム
テキスト
出現情報ファイル
出現確率ファイル
出現回数ファイル
共有ライブラリ
単一文字種文字列抽出プログラム
ワークエリア

【図3】

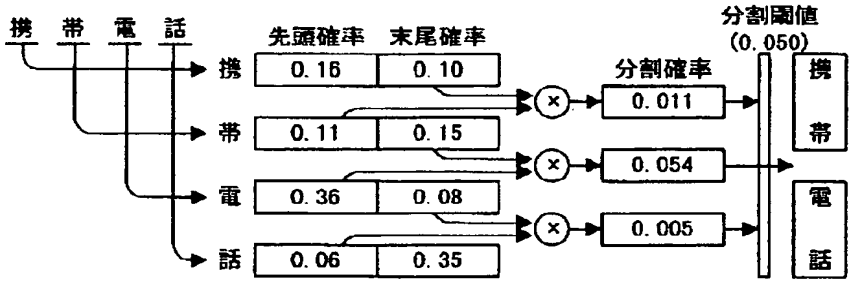
図3

No.	1-gram	先頭確率	末尾確率
1	一	0.00	0.22
2	ナ	0.09	0.07
3	マ	0.35	0.14
4	携	0.16	0.10
5	帯	0.11	0.15
6	題	0.02	0.89
7	電	0.36	0.08
8	問	0.42	0.26
9	用	0.09	0.56
10	話	0.06	0.34

300

【図4】

図4



【図5】

図5

No.	n-gram	出現回数	先頭回数	末尾回数	単独回数
1	一	62,318	0	13,480	0
2	ナ	28,090	2,653	2,079	0
3	マ	43,300	15,235	6,179	0
4	携	4,740	768	492	42
5	帯	4,703	530	687	26
6	題	36,338	733	32,342	332
7	電	38,317	13,794	3,218	218
8	問	46,216	19,205	11,884	884
9	用	59,987	5,132	33,600	768
10	話	18,416	1,105	6,353	211
11	ナ一	3,867	65	3,040	0
12	マナ	122	99	4	0
13	携帯	462	419	52	48
14	使用	2,704	2,156	1,517	517
15	帯電	269	14	4	4
16	電話	5,283	2,538	3,053	1,298
17	問題	29,095	15,280	25,547	13,157

500

【図6】

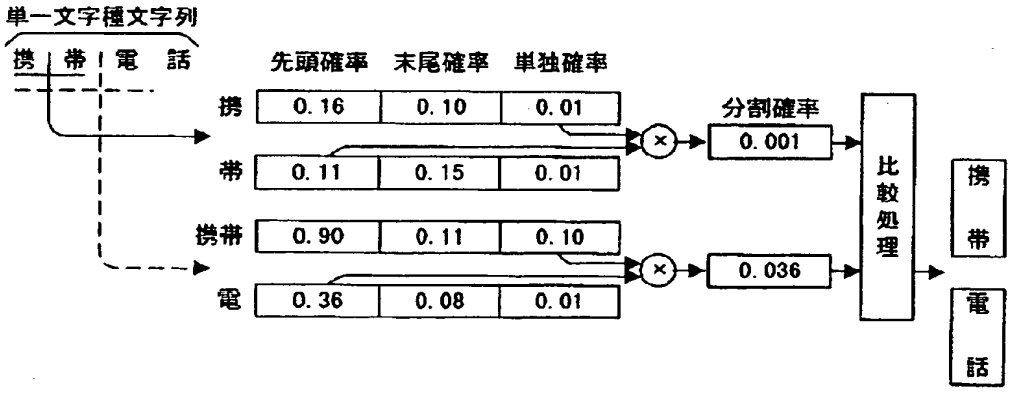
図6

No.	n-gram	先頭確率	末尾確率	単独確率
1	一	0.00	0.22	0.00
2	ナ	0.09	0.07	0.00
3	マ	0.35	0.14	0.00
4	携	0.16	0.10	0.01
5	使	0.69	0.70	0.12
6	帯	0.11	0.15	0.01
7	題	0.02	0.89	0.01
8	電	0.36	0.08	0.01
9	問	0.42	0.26	0.02
10	用	0.09	0.56	0.01
11	話	0.06	0.34	0.01
12	ナー	0.02	0.79	0.00
13	マナ	0.81	0.03	0.00
14	携帯	0.90	0.11	0.10
15	使用	0.80	0.56	0.19
16	帯電	0.05	0.01	0.01
17	電話	0.48	0.58	0.25
18	問題	0.53	0.88	0.45

600

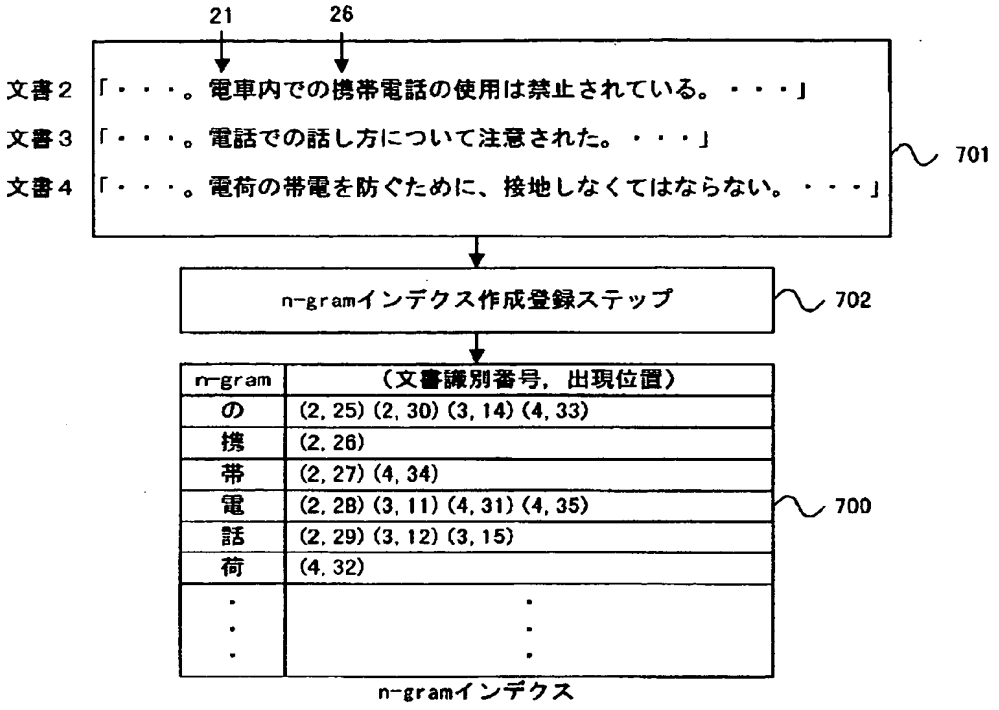
【図8】

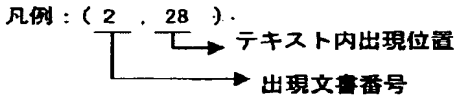
図8



【図7】

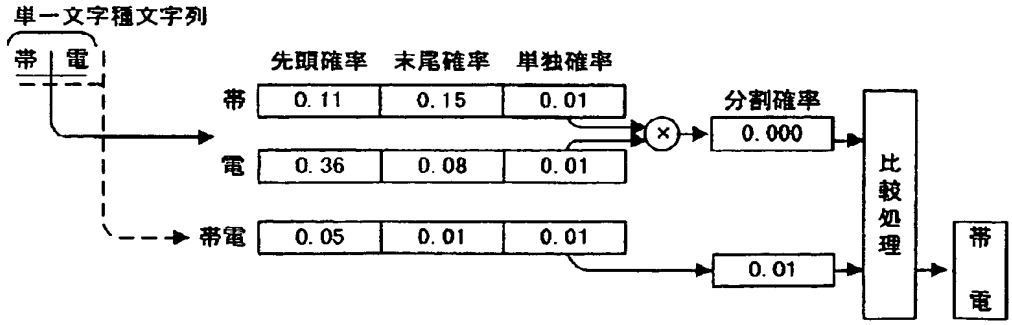
図7


21
26
文書2「・・・。電車内での携帯電話の使用は禁止されている。・・・」
文書3「・・・。電話での話し方について注意された。・・・」
文書4「・・・。電荷の帯電を防ぐために、接地しなくてはならない。・・・」
701
n-gramインデクス作成登録ステップ
702
n-gram
（文書識別番号，出現位置）
の
(2, 25) (2, 30) (3, 14) (4, 33)
携
(2, 26)
帯
(2, 27) (4, 34)
電
(2, 28) (3, 11) (4, 31) (4, 35)
話
(2, 29) (3, 12) (3, 15)
荷
(4, 32)
700
n-gramインデクス


凡例：(2 , 28)
テキスト内出現位置
出現文書番号

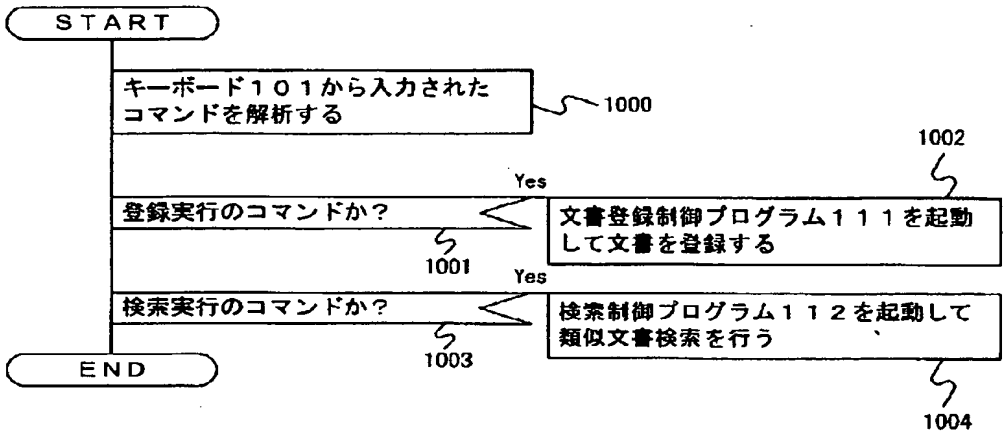
【図9】

図9


単一文字種文字列
帯
電
先頭確率
末尾確率
単独確率
帯
0.11
0.15
0.01
電
0.36
0.08
0.01
帯電
0.05
0.01
0.01
分割確率
0.000
0.01
比較処理
帯
電

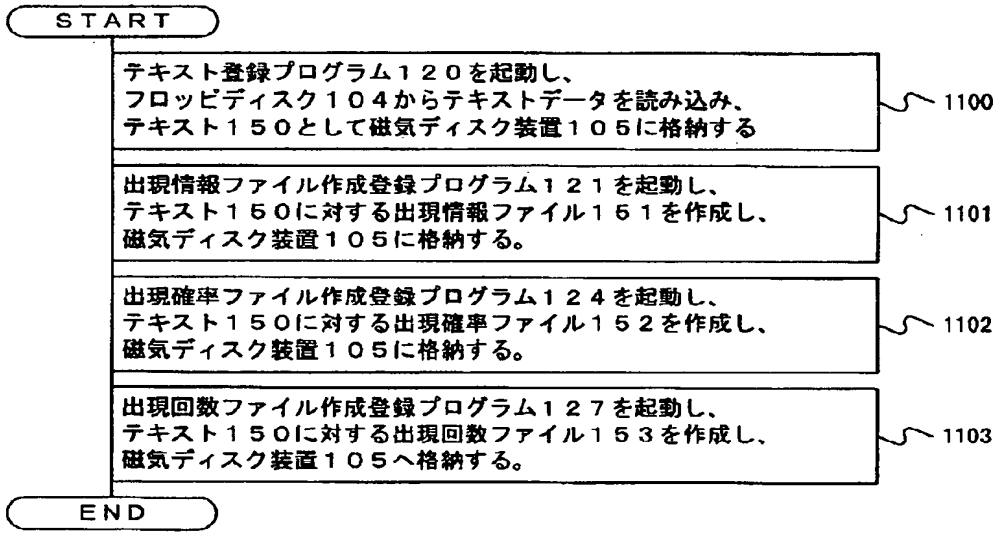
【図10】

図10



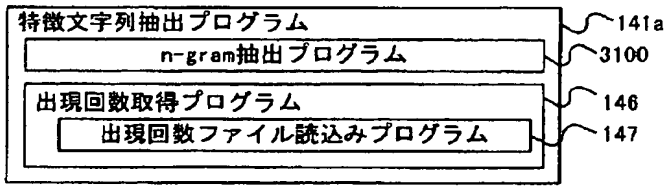
【図11】

図11



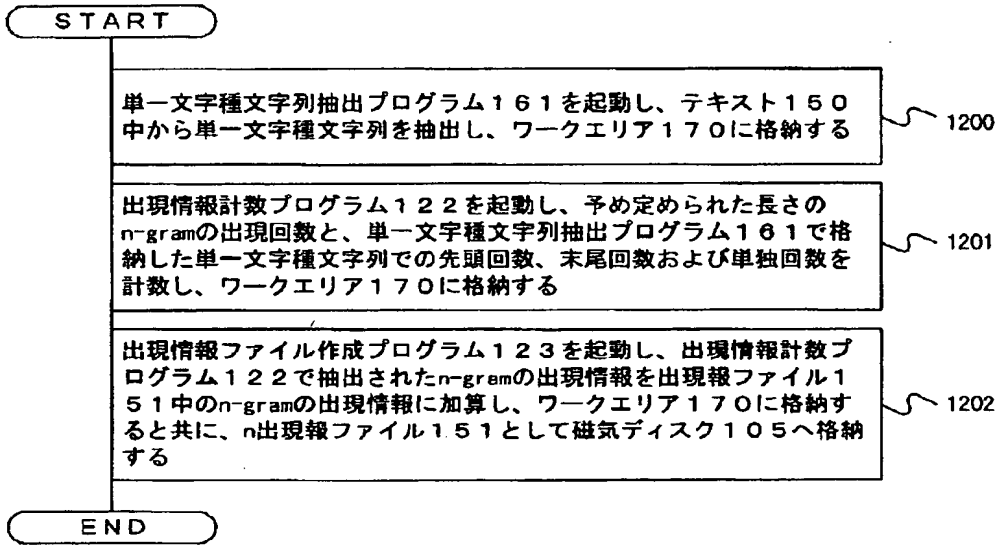
【図31】

図31



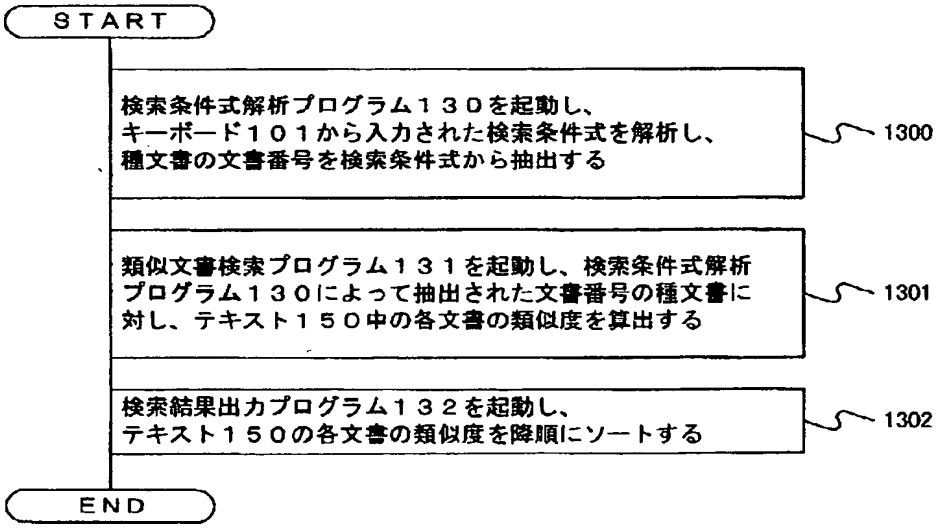
【図12】

図12



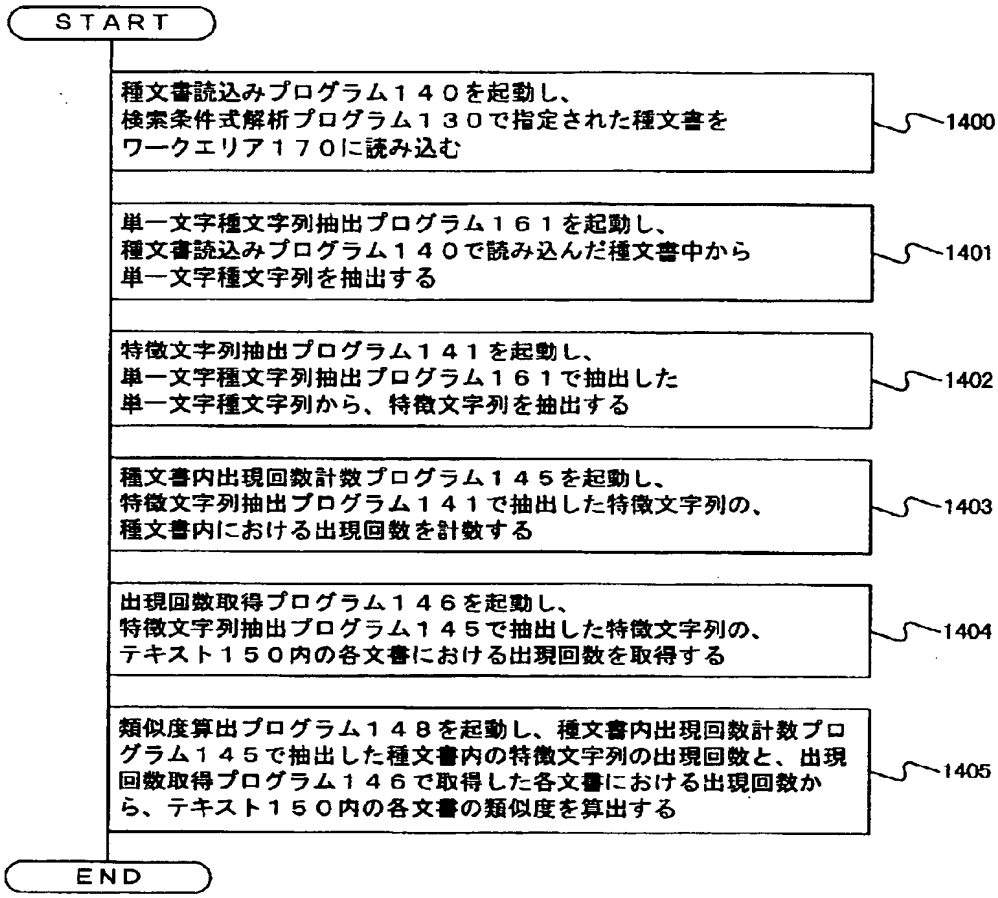
【図13】

図13



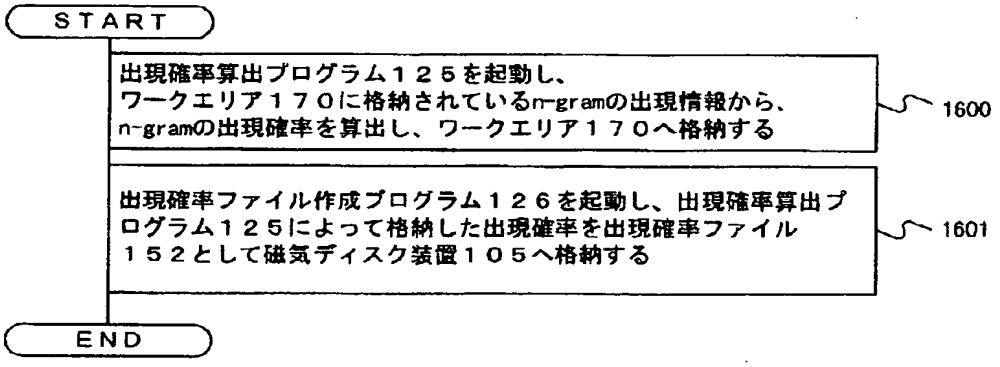
【図14】

図14



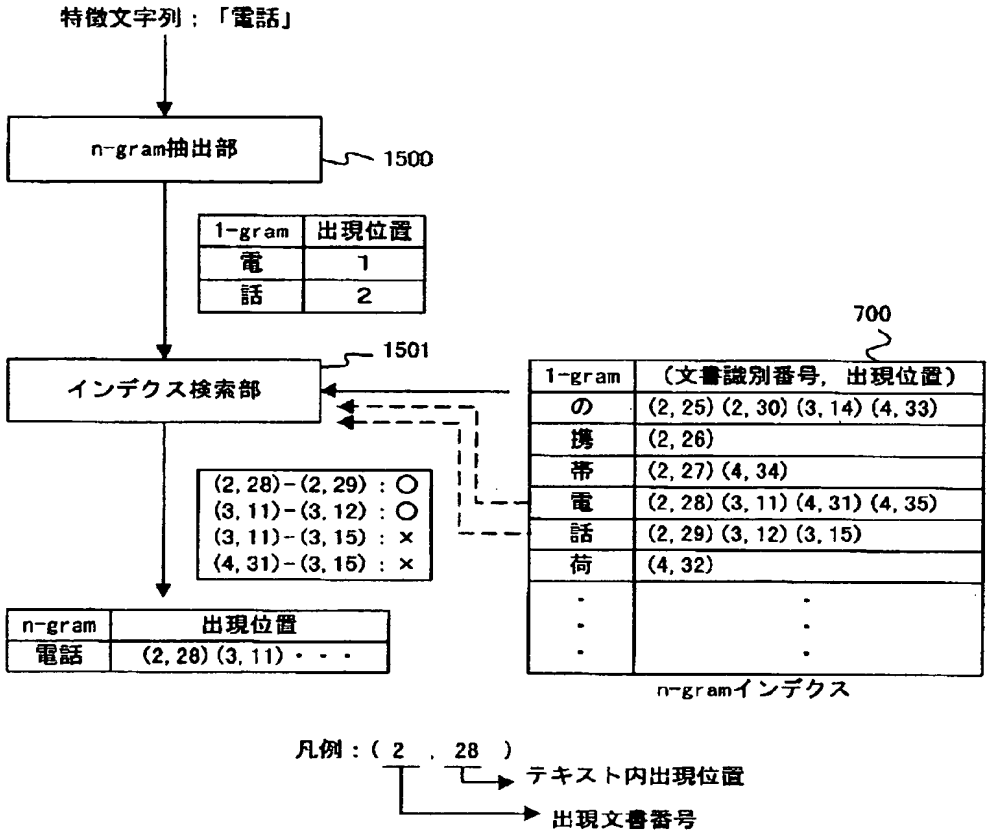
【図16】

図16



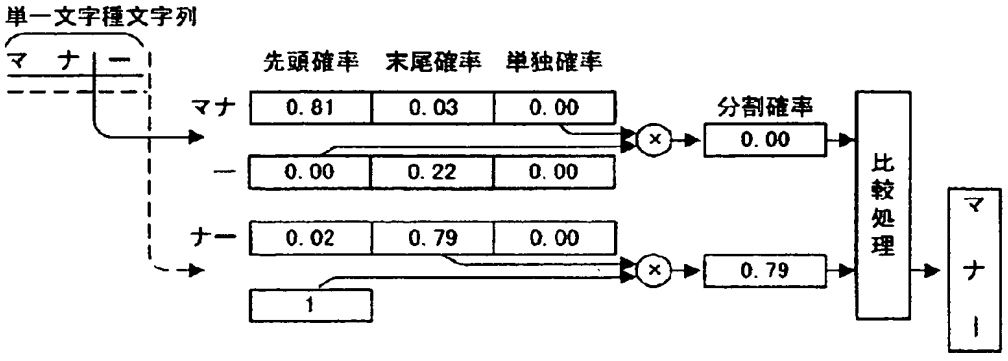
【図１５】

図１５



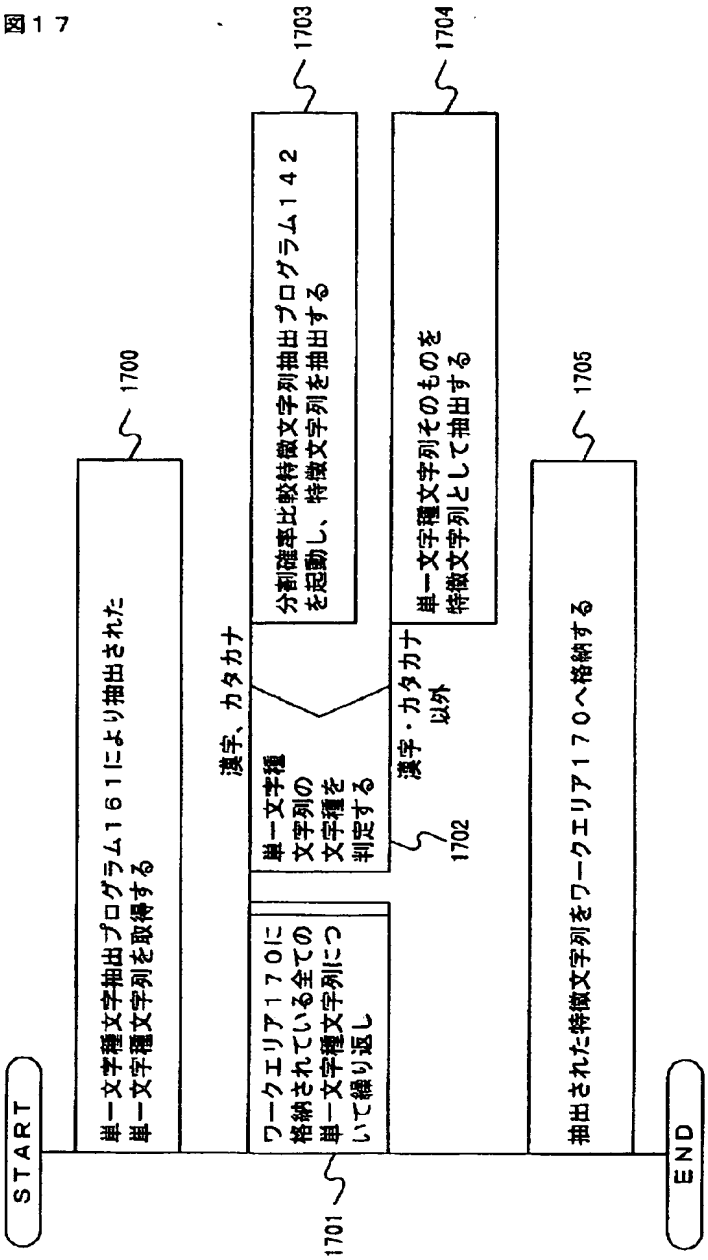
【図２０】

図２０



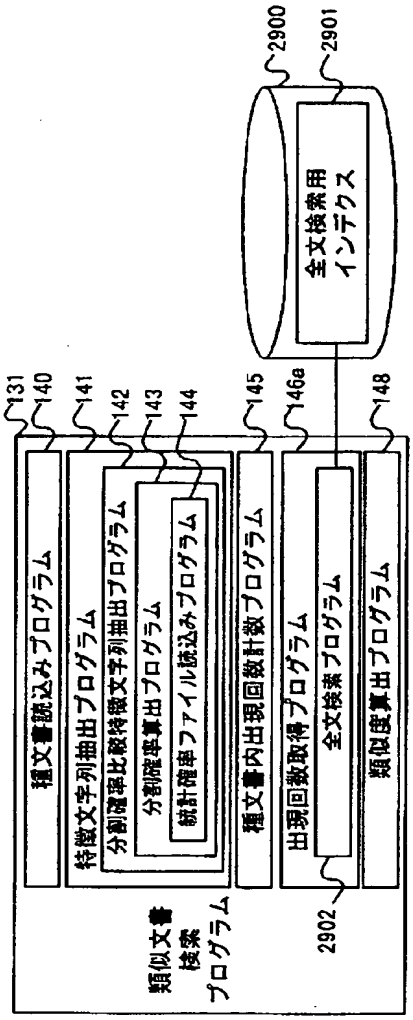
【図17】

図17



【図29】

図29



【図18】

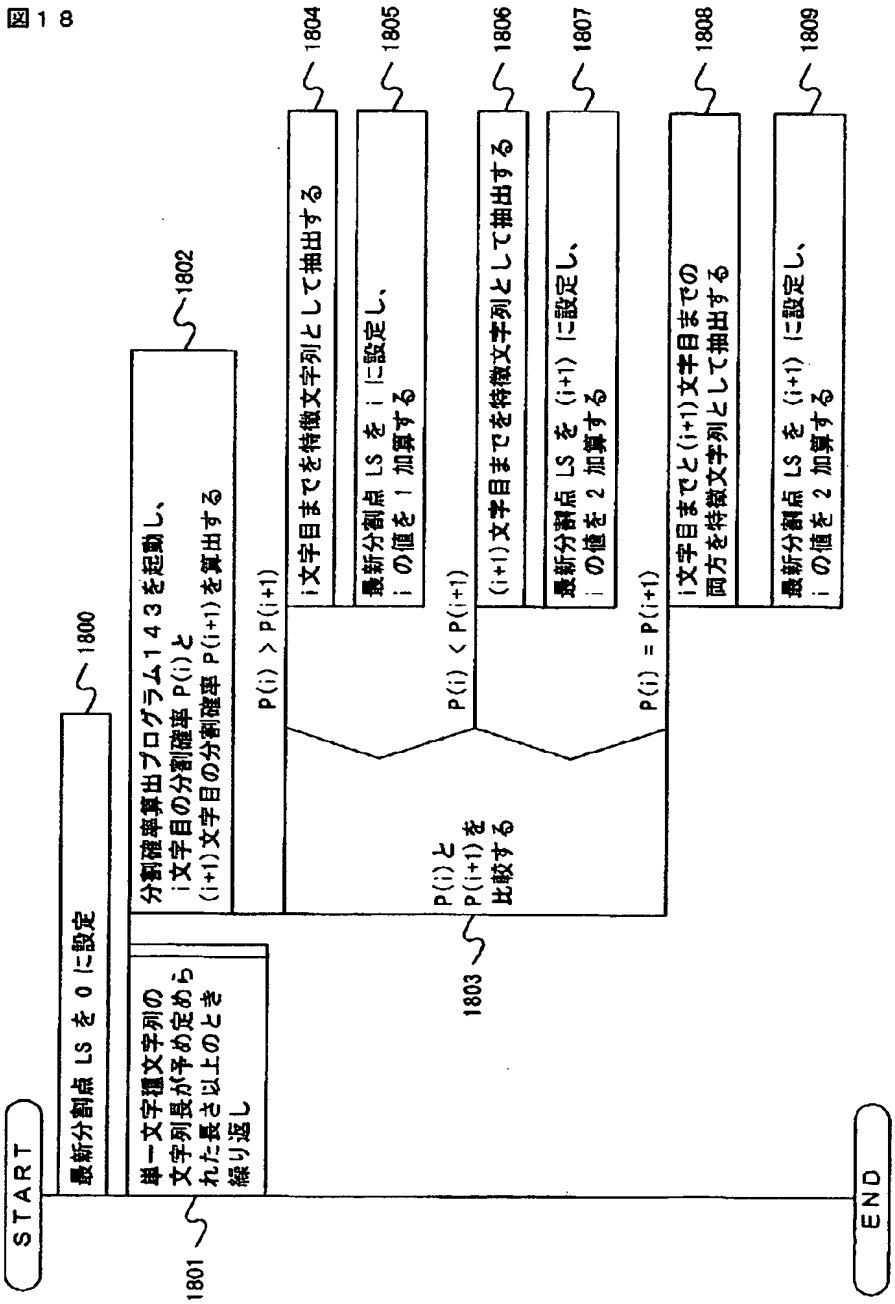
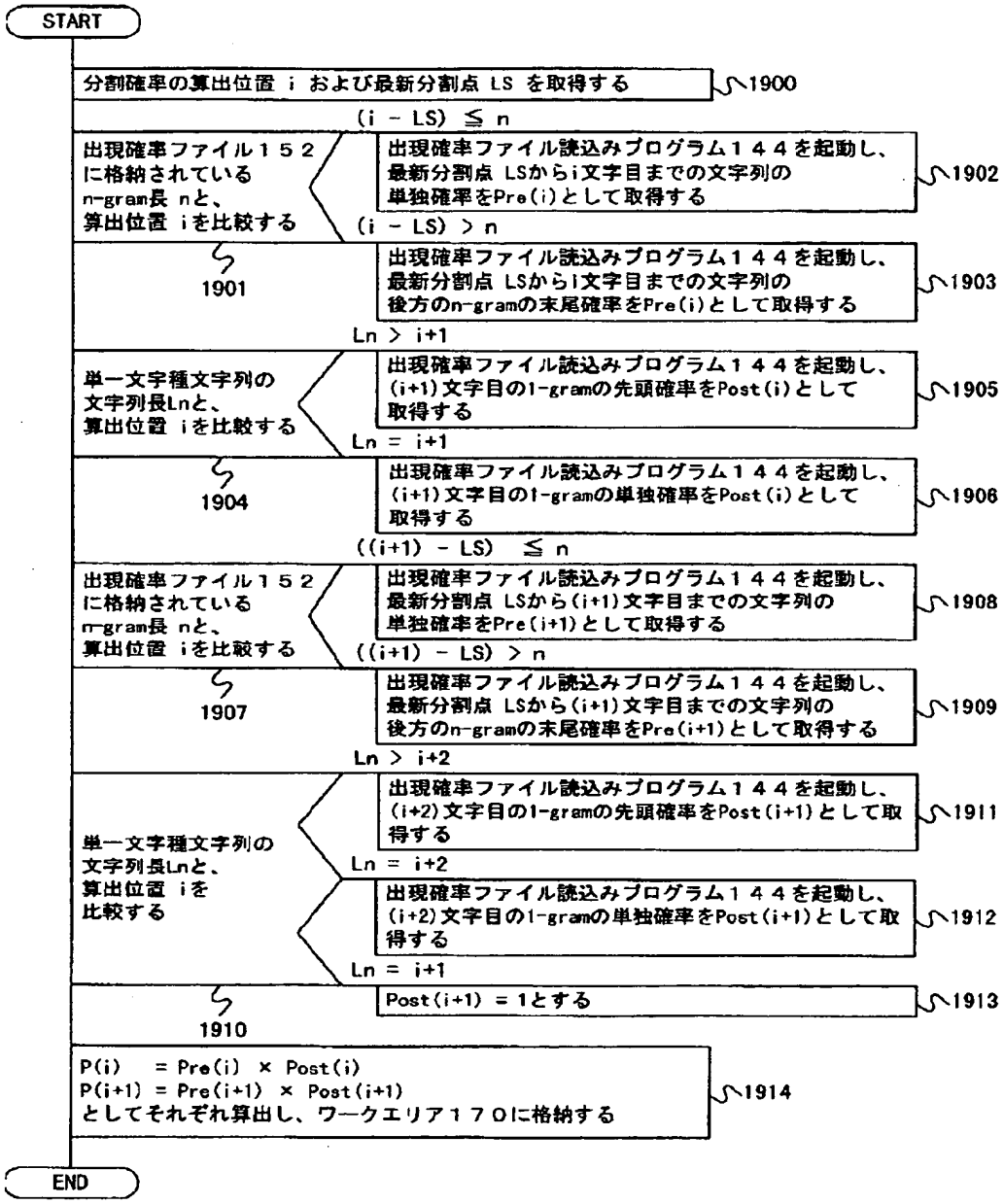


図18
START
最新分割点 LS を 0 に設定
1800
単一文字種文字列の
文字列長が予め定められた長さ以上のとき
繰り返し
1801
分割確率算出プログラム143を起動し、
i文字目の分割確率 P(i)と
(i+1)文字目の分割確率 P(i+1)を算出する
1802
P(i)と
P(i+1)を
比較する
1803
P(i) > P(i+1)
i文字目までを特徴文字列として抽出する
1804
最新分割点 LS を i に設定し、
i の値を 1 加算する
1805
P(i) < P(i+1)
(i+1)文字目までを特徴文字列として抽出する
1806
最新分割点 LS を (i+1) に設定し、
i の値を 2 加算する
1807
P(i) = P(i+1)
i文字目までと(i+1)文字目までの
両方を特徴文字列として抽出する
1808
最新分割点 LS を (i+1) に設定し、
i の値を 2 加算する
1809
END

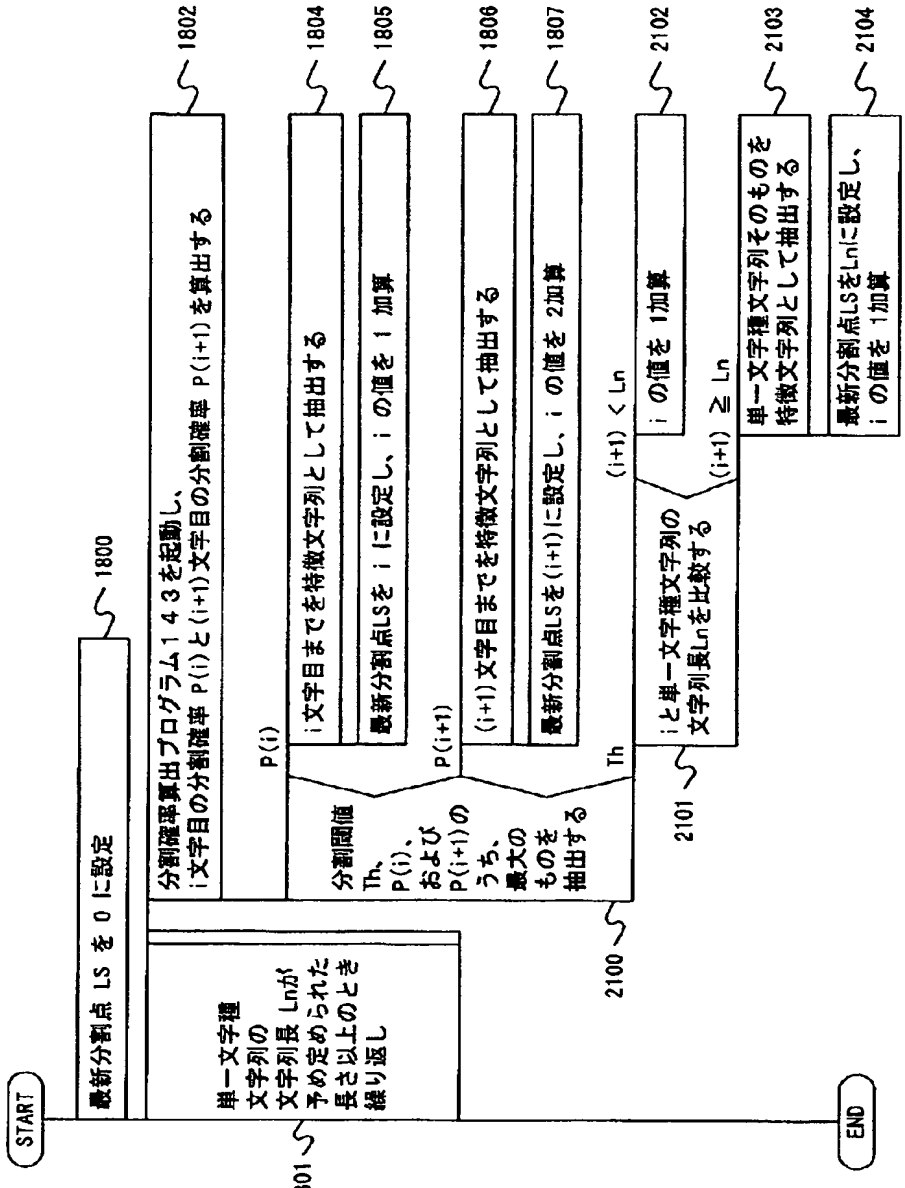
【図１９】

図１９


START
分割確率の算出位置 i および最新分割点 LS を取得する
1900
(i - LS) ≦ n
出現確率ファイル１５２に格納されているn-gram長 nと、算出位置 iを比較する
出現確率ファイル読込みプログラム１４４を起動し、最新分割点 LSからi文字目までの文字列の単独確率をPre(i)として取得する
1902
(i - LS) > n
1901
出現確率ファイル読込みプログラム１４４を起動し、最新分割点 LSからi文字目までの文字列の後方のn-gramの末尾確率をPre(i)として取得する
1903
Ln > i+1
単一文字種文字列の文字列長Lnと、算出位置 iを比較する
出現確率ファイル読込みプログラム１４４を起動し、(i+1)文字目の1-gramの先頭確率をPost(i)として取得する
1905
Ln = i+1
1904
出現確率ファイル読込みプログラム１４４を起動し、(i+1)文字目の1-gramの単独確率をPost(i)として取得する
1906
((i+1) - LS) ≦ n
出現確率ファイル１５２に格納されているn-gram長 nと、算出位置 iを比較する
出現確率ファイル読込みプログラム１４４を起動し、最新分割点 LSから(i+1)文字目までの文字列の単独確率をPre(i+1)として取得する
1908
((i+1) - LS) > n
1907
出現確率ファイル読込みプログラム１４４を起動し、最新分割点 LSから(i+1)文字目までの文字列の後方のn-gramの末尾確率をPre(i+1)として取得する
1909
Ln > i+2
単一文字種文字列の文字列長Lnと、算出位置 iを比較する
出現確率ファイル読込みプログラム１４４を起動し、(i+2)文字目の1-gramの先頭確率をPost(i+1)として取得する
1911
Ln = i+2
出現確率ファイル読込みプログラム１４４を起動し、(i+2)文字目の1-gramの単独確率をPost(i+1)として取得する
1912
Ln = i+1
1910
Post(i+1) = 1とする
1913
P(i) = Pre(i) × Post(i)
P(i+1) = Pre(i+1) × Post(i+1)
としてそれぞれ算出し、ワークエリア１７０に格納する
1914
END

【図21】

図21


START
最新分割点 LS を 0 に設定
1800
単一文字種
文字列の
文字列長 Lnが
予め定められた
長さ以上のとき
繰り返し
1801
分割確率算出プログラム143を起動し、
i文字目の分割確率 P(i)とi+1)文字目の分割確率 P(i+1)を算出する
1802
分割閾値
Th、
P(i)、
および
P(i+1)の
うち、
最大の
ものを
抽出する
2100
P(i)
i文字目までを特徴文字列として抽出する
1804
最新分割点LSを i に設定し、i の値を 1 加算
1805
P(i+1)
(i+1)文字目までを特徴文字列として抽出する
1806
最新分割点LSを(i+1)に設定し、i の値を 2加算
1807
Th
(i+1) < Ln
i の値を 1加算
2102
2101
iと単一文字種文字列の
文字列長Lnを比較する
(i+1) ≧ Ln
単一文字種文字列そのものを
特徴文字列として抽出する
2103
最新分割点LSをLnに設定し、
i の値を 1加算
2104
END

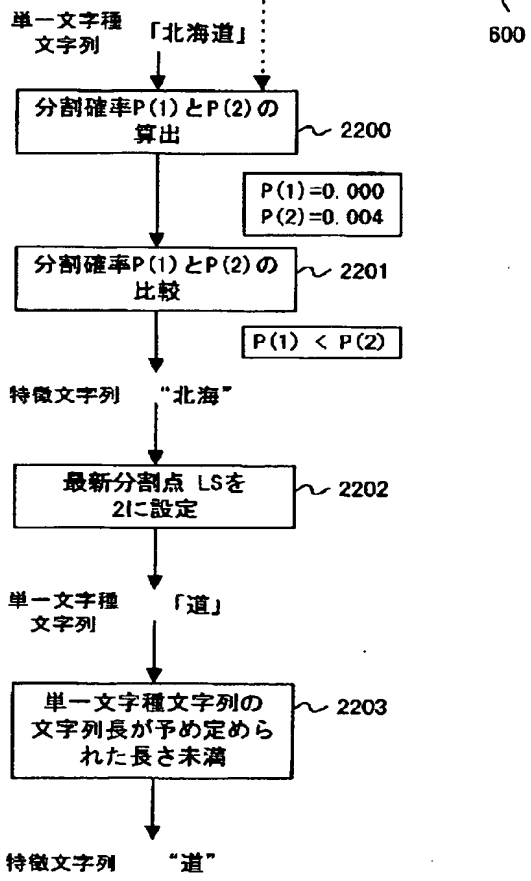
【図２２】

図２２

出現確率ファイル１５２

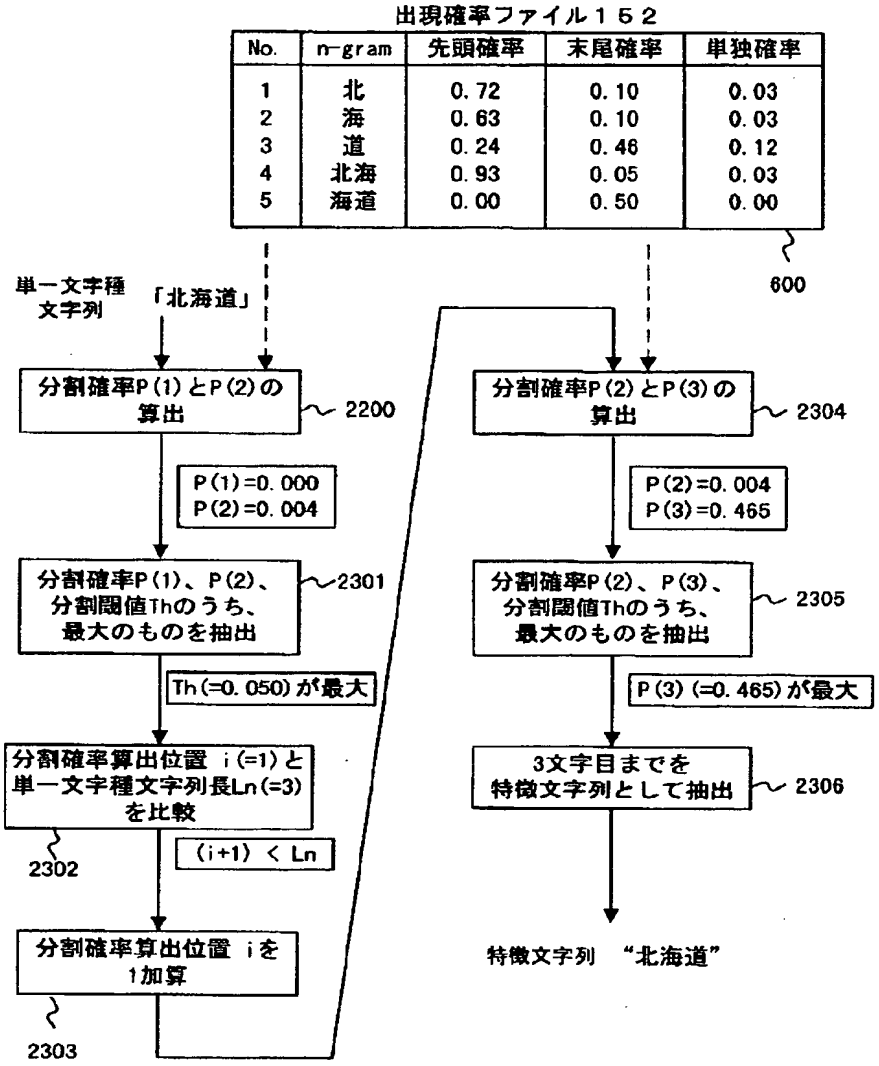
No.	n-gram	先頭確率	末尾確率	単独確率
1	北	0.72	0.10	0.03
2	海	0.63	0.10	0.03
3	道	0.24	0.46	0.12
4	北海	0.93	0.05	0.03
5	海道	0.00	0.50	0.00

600



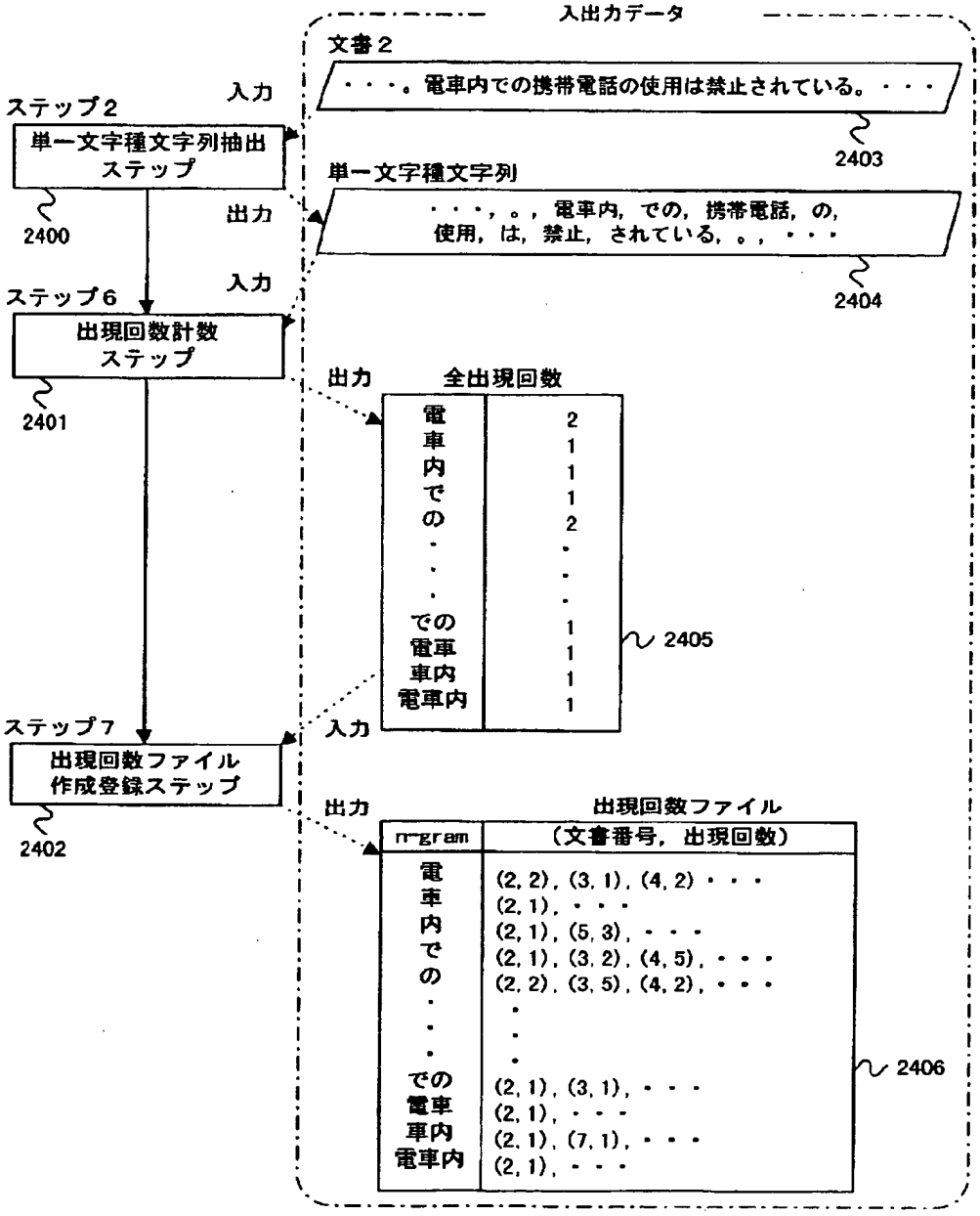
【図２３】

図２３


出現確率ファイル１５２
No. | n-gram | 先頭確率 | 末尾確率 | 単独確率
1 | 北 | 0.72 | 0.10 | 0.03
2 | 海 | 0.63 | 0.10 | 0.03
3 | 道 | 0.24 | 0.46 | 0.12
4 | 北海 | 0.93 | 0.05 | 0.03
5 | 海道 | 0.00 | 0.50 | 0.00
600
単一文字種文字列
「北海道」
分割確率P(1)とP(2)の算出
2200
P(1)=0.000
P(2)=0.004
分割確率P(1)、P(2)、分割閾値Thのうち、最大のものを抽出
2301
Th(=0.050)が最大
分割確率算出位置 i(=1)と単一文字種文字列長Ln(=3)を比較
2302
(i+1) < Ln
分割確率算出位置 iを1加算
2303
分割確率P(2)とP(3)の算出
2304
P(2)=0.004
P(3)=0.465
分割確率P(2)、P(3)、分割閾値Thのうち、最大のものを抽出
2305
P(3) (=0.465)が最大
3文字目までを特徴文字列として抽出
2306
特徴文字列　“北海道”

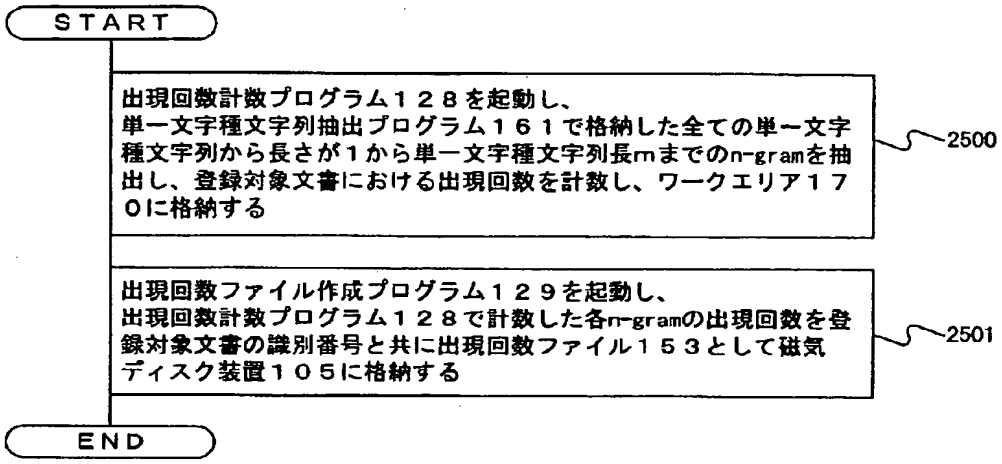
【図２４】

図２４


入出力データ
文書２
・・・。電車内での携帯電話の使用は禁止されている。・・・
2403
ステップ2
入力
単一文字種文字列抽出ステップ
2400
出力
単一文字種文字列
・・・，。，電車内，での，携帯電話，の，使用，は，禁止，されている，。，・・・
2404
ステップ6
入力
出現回数計数ステップ
2401
出力
全出現回数
電車内での・・・での電車車内電車内
2 1 1 1 2 ・ ・ ・ 1 1 1 1
2405
ステップ7
入力
出現回数ファイル作成登録ステップ
2402
出力
出現回数ファイル
n-gram
（文書番号，出現回数）
電車内での・・・での電車車内電車内
(2, 2), (3, 1), (4, 2) ・・・
(2, 1), ・・・
(2, 1), (5, 3), ・・・
(2, 1), (3, 2), (4, 5), ・・・
(2, 2), (3, 5), (4, 2), ・・・
(2, 1), (3, 1), ・・・
(2, 1), ・・・
(2, 1), (7, 1), ・・・
(2, 1), ・・・
2406

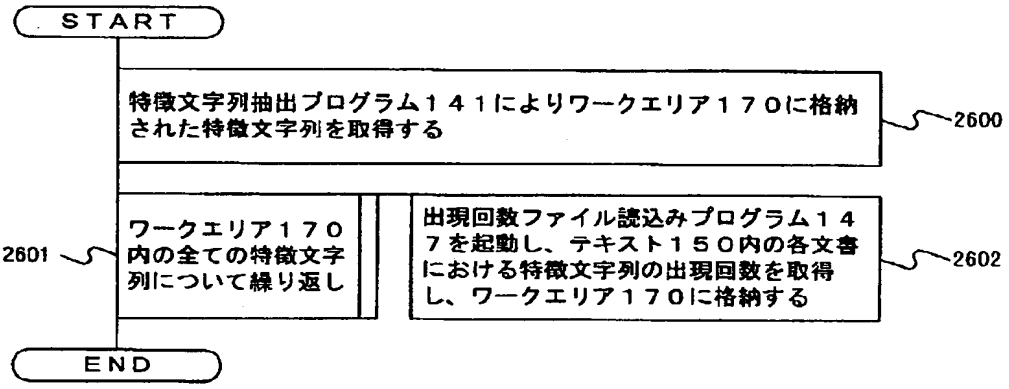
【図25】

図25



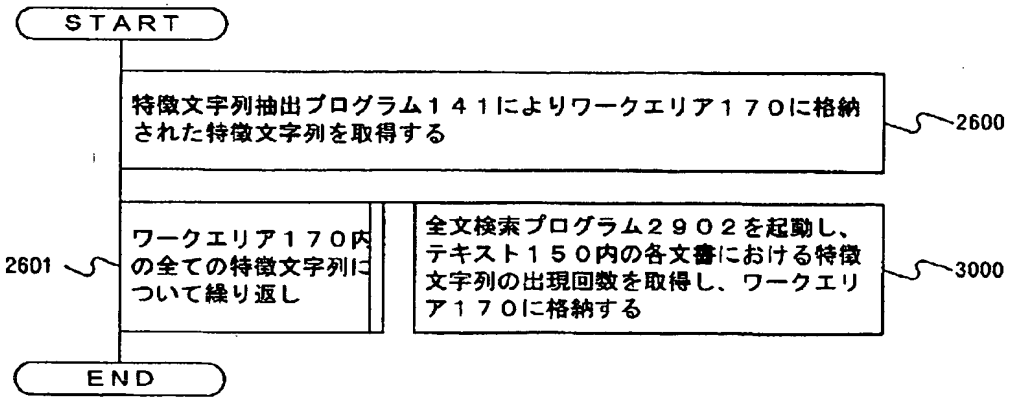
【図26】

図26



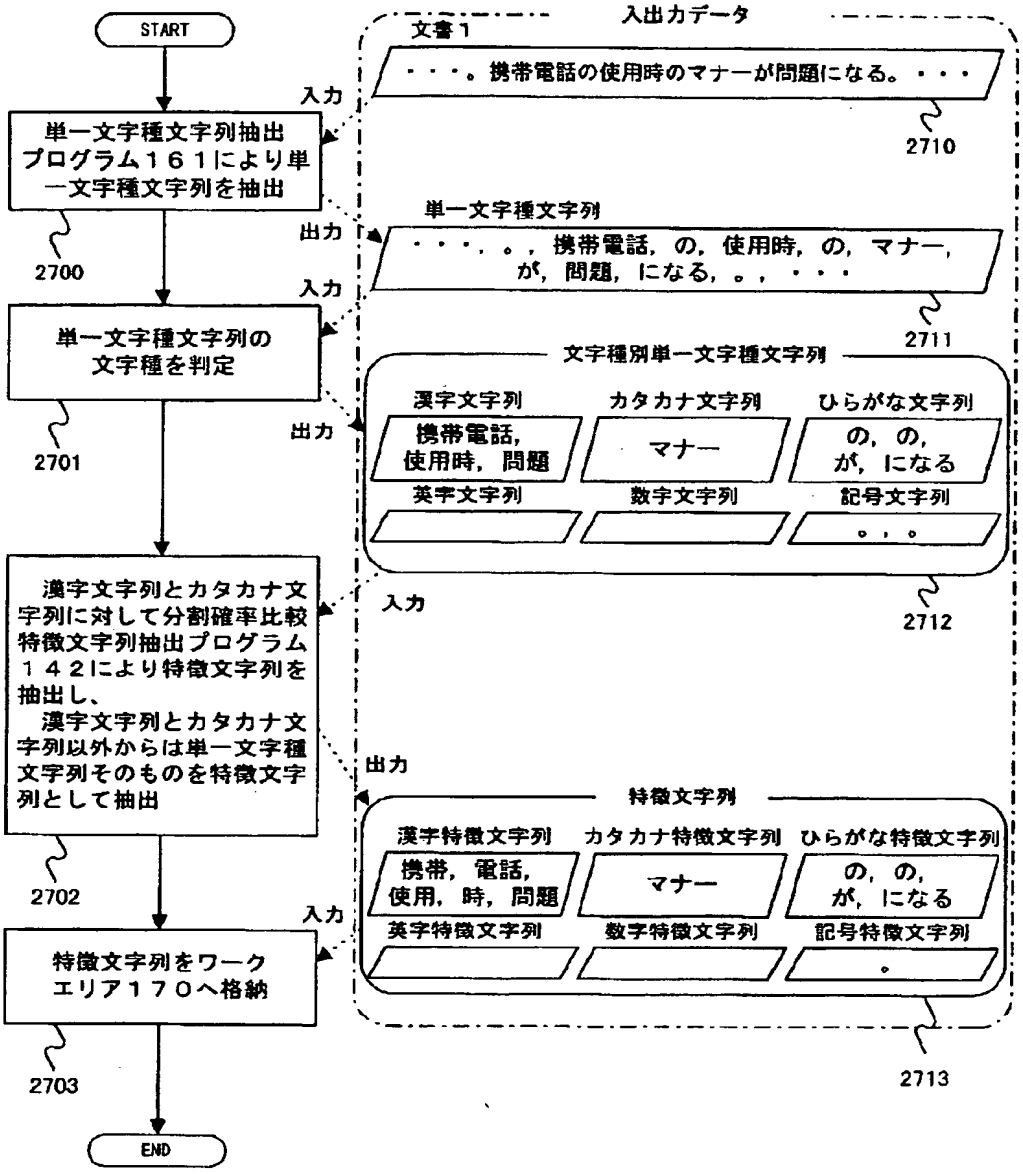
【図30】

図30



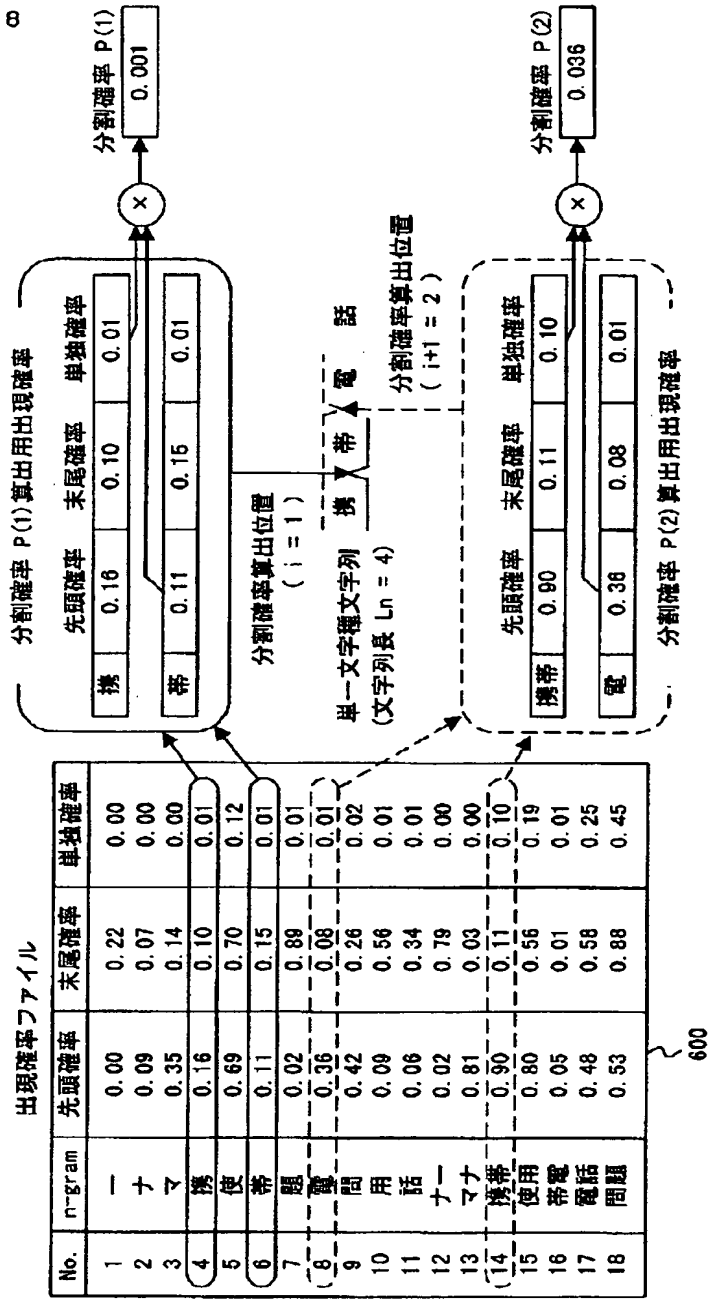
【図27】

図27


START
単一文字種文字列抽出プログラム161により単一文字種文字列を抽出
2700
単一文字種文字列の文字種を判定
2701
漢字文字列とカタカナ文字列に対して分割確率比較特徴文字列抽出プログラム142により特徴文字列を抽出し、
漢字文字列とカタカナ文字列以外からは単一文字種文字列そのものを特徴文字列として抽出
2702
特徴文字列をワークエリア170へ格納
2703
END
入力
出力
入出力データ
文書1
・・・。携帯電話の使用時のマナーが問題になる。・・・
2710
単一文字種文字列
・・・。。携帯電話，の，使用時，の，マナー，が，問題，になる，。，・・・
2711
文字種別単一文字種文字列
漢字文字列
携帯電話，使用時，問題
カタカナ文字列
マナー
ひらがな文字列
の，の，が，になる
英字文字列
数字文字列
記号文字列
。，。
2712
特徴文字列
漢字特徴文字列
携帯，電話，使用，時，問題
カタカナ特徴文字列
マナー
ひらがな特徴文字列
の，の，が，になる
英字特徴文字列
数字特徴文字列
記号特徴文字列
。
2713

【図28】

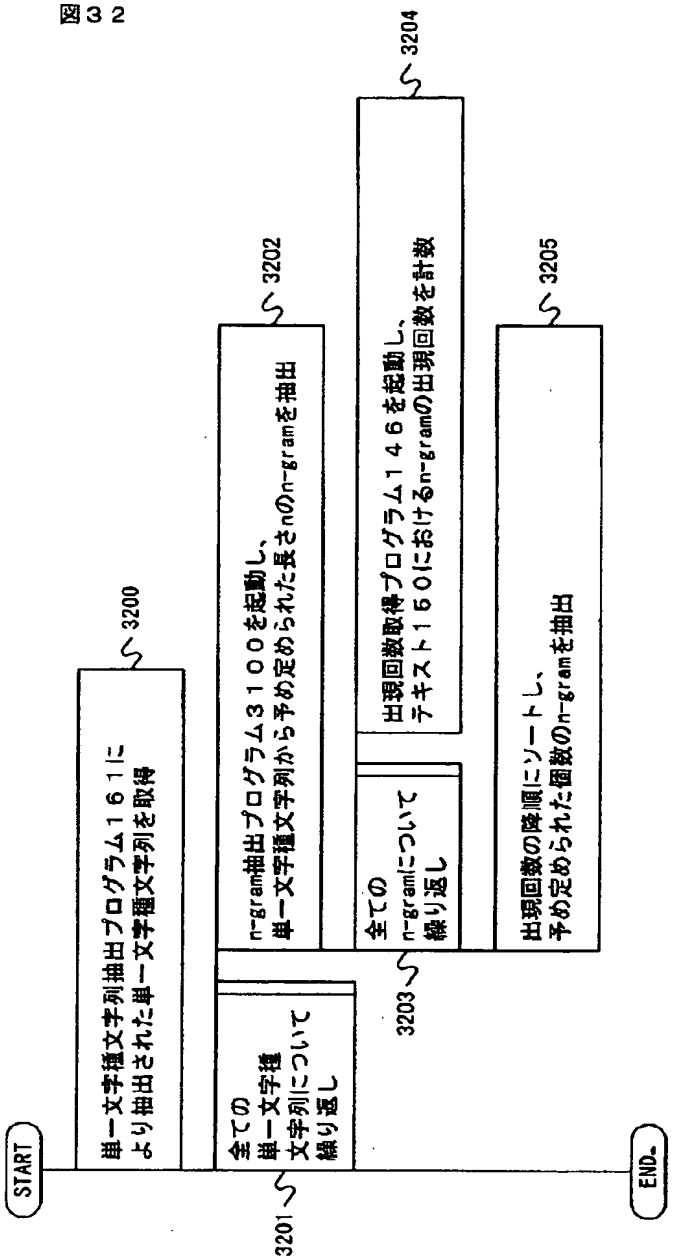
図28



No.	n-gram	先頭確率	末尾確率	単独確率
1	一	0.00	0.22	0.00
2	十	0.09	0.07	0.00
3	マ	0.35	0.14	0.00
4	携	0.16	0.10	0.01
5	使	0.69	0.70	0.12
6	帯	0.11	0.15	0.01
7	題	0.02	0.89	0.01
8	電	0.36	0.08	0.01
9	間	0.42	0.26	0.02
10	用	0.09	0.56	0.01
11	話	0.06	0.34	0.01
12	ナー	0.02	0.79	0.00
13	マナ	0.81	0.03	0.00
14	携帯	0.90	0.11	0.10
15	使用	0.80	0.56	0.19
16	帯電	0.05	0.01	0.01
17	電話	0.48	0.58	0.25
18	問題	0.53	0.88	0.45

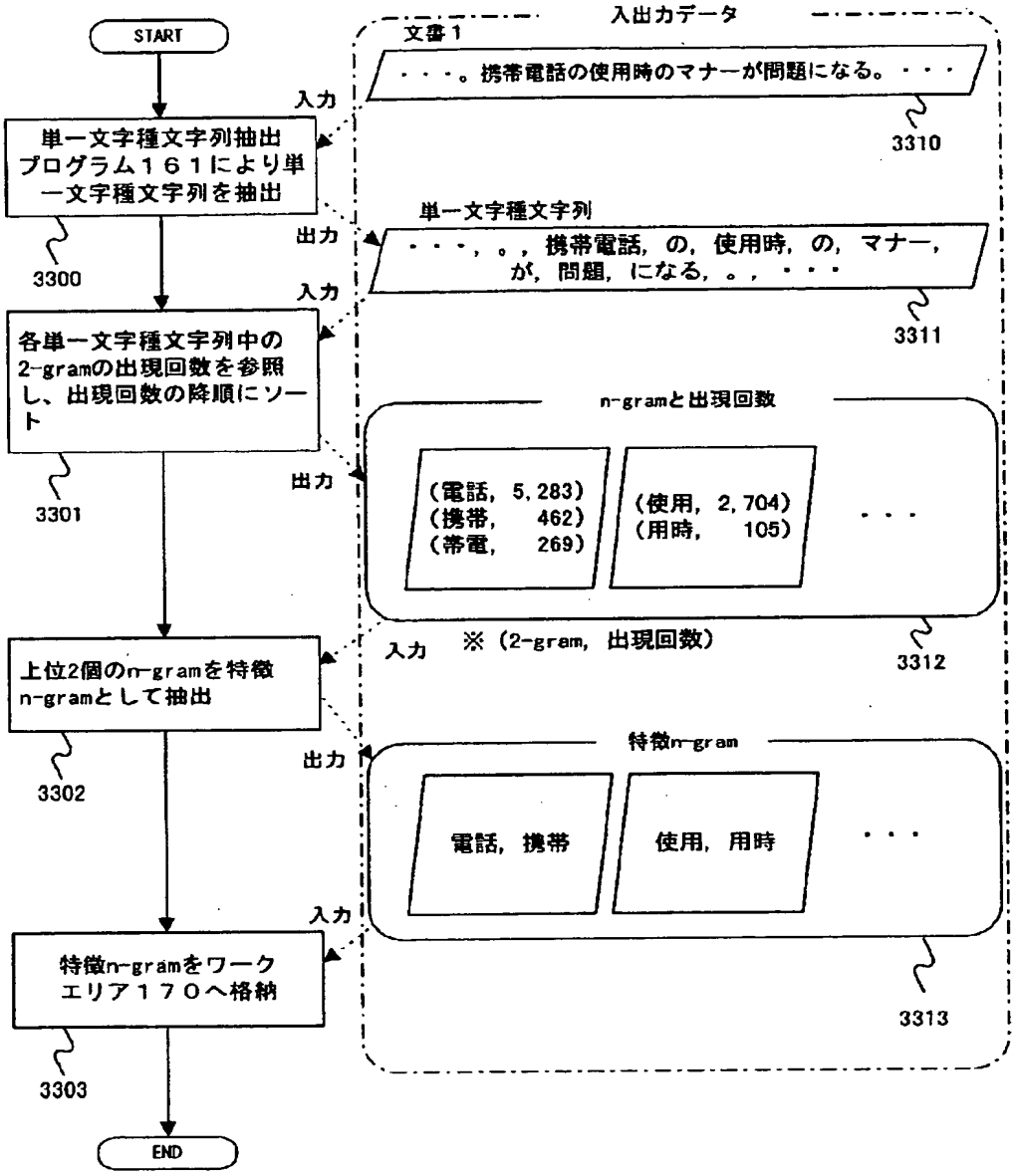
【図32】

図32


START
END
3200
単一文字種文字列抽出プログラム161により抽出された単一文字種文字列を取得
3201
全ての単一文字種文字列について繰り返し
3202
n-gram抽出プログラム3100を起動し、単一文字種文字列から予め定められた長さnのn-gramを抽出
3203
全てのn-gramについて繰り返し
3204
出現回数取得プログラム146を起動し、テキスト150におけるn-gramの出現回数を計数
3205
出現回数の降順にソートし、予め定められた個数のn-gramを抽出

【図３３】

図３３



フロントページの続き

(72)発明者　菅谷　奈津子
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内

(72)発明者　川下　靖司
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内